NBER WORKING PAPER SERIES

THE CYCLICAL BEHAVIOR OF PRICES AND COSTS

Julio J. Rotemberg
Michael Woodford

Working Paper 6909
http://www.nber.org/papers/w6909

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 1999

Prepared for John B. Taylor and Michael Woodford, eds., *Handbook of Macroeconomics*, North-Holland, forthcoming. We wish to thank Mark Bils, Susanto Basu, Robert Chirinko, Miles Kimball, Argia Sbordone, and John Taylor for comments. We are also grateful for research support from the Harvard Business School Division of Research and from the NSF through a grant to the NBER. The views expressed here are those of the author and do not reflect those of the National Bureau of Economic Research.

The Cyclical Behavior of Prices and Costs
Julio J. Rotemberg and Michael Woodford
NBER Working Paper No. 6909
January 1999
JEL No. E32

ABSTRACT

Because inputs are scarce, marginal cost should be an increasing function of output. Without changes in this *real marginal cost schedule*, aggregate output can vary if and only if the markup of price over marginal cost varies. In this review, we discuss the extent to which observed fluctuations in aggregate economic activity depend upon such variations in average markups.

We first study whether, empirically, real marginal cost rises in cyclical expansions. Average real labor cost is not very procyclical, but, for reasons such as overhead labor and adjustment costs, marginal labor cost should be more procyclical. Measures of marginal cost based on materials costs and inventories also appear procyclical.

We next show that countercyclical markup variation may, depending upon how costs are modeled, account for a substantial fraction of cyclical output movements. We also show that the observed procyclical variations in productivity and profits are consistent with the hypothesis that cyclical variations in output are primarily due to markup variations than to shifts in the real marginal cost schedule.

Finally, we survey theories of endogenous markup variation. These include both models of sticky and models in which firms' desired markup varies over time.

Julio J. Rotemberg
Harvard Business School
Soldiers Field Road
Boston, MA 02142
andNBER
jrotemberg@hbs.edu

Michael Woodford
Department of Economics
Princeton University
Princeton, NJ 80544
and NBER
woodford@princeton.edu

1 Introduction: Markups and the Business Cycle

In this paper, we consider the role of variations in the relationship between the prices at which goods are supplied and the marginal cost of supplying them in accounting for observed fluctuations in economic activity and employment. We shall argue that there exists a great deal of evidence in support of the view that marginal costs rise more than prices in economic expansions, especially late in expansions. Thus *real marginal cost* (MC/P) rises, for the typical firm; and alternatively, the *markup* of price over marginal cost (which we shall define as the ratio P/MC[1]) declines for the typical firm. These two ways of describing the feature of business cycles with which we are concerned are equivalent in the case of a symmetric equilibrium (in which the costs, output and prices of all goods move exactly together), though they are not equivalent propositions regarding an individual firm or industry in the asymmetric case (since the firm's or industry's relative price may vary). Because of our concern with the explanation of fluctuations in aggregate activity, and because such aggregate fluctuations are characterized by a striking degree of comovement among sectors, we will mainly conduct our analysis in terms of a symmetric (aggregative) model, and treat procyclical movements in real marginal cost as equivalent to countercyclical markup variations. Discussion of the cyclicality of real marginal cost is most natural when one is discussing measurement (since the crucial measurement issue is to infer the level of marginal costs), and so this is how the issue is framed in many empirical studies (such as Bils, 1987, and Bils and Kahn, 1996). But when we turn to theoretical explanation, it is most useful to describe the phenomenon in terms of variation in firms' markups, because the crucial decisions responsible for the phenomenon are firms' decisions about the prices at which they are willing to supply their products. In fact, in a broad class of models discussed below, the cyclicality of real marginal costs in equilibrium turns not upon the nature of the production technology or the conditions under which firms can obtain factors of production, but upon the nature of the competition among firms in their product markets; and so, from an analytical point of view, it seems most important to emphasize variation in markups.

The observation that costs rise more than do prices, at least late in expansions, is not a new one. It was emphasized in the work of Wesley Clair Mitchell, for example, who writes (1941, p. 52) that as activity expands, "equipment of less than standard efficiency is brought back into use; the price of labor rises while its efficiency falls; the cost of materials, supplies and wares for resale advances faster than selling prices; discount rates go up at an especially rapid pace, and all the little wastes incidental to the conduct of business enterprises grow steadily larger." That real marginal costs should rise for such reasons is, of course, a simple consequence of the fact that factors of production are not in unlimited supply. But, as Mitchell notes, from the point of view of theory "a problem still remains: Why cannot businessmen defend their profit margins against the threatened encroachment of costs by marking up their selling prices?" As we shall see, a variety of theories of imperfectly competitive behavior by firms explain why they may choose not to do so. Such imperfectly competitive behavior thus plays an important role in accounting for the character of aggregate fluctuations.

Fluctuations in markups are an important factor, in our view, for a reason somewhat different than that emphasized by Mitchell. In Mitchell's analysis, the squeezing of profit margins late in booms is what brings the boom to an end, as reduced profitability dampens investment demand and hence sales. This suggests that an improvement in firms' power to set prices above marginal cost would extend the boom. But this neglects the fact that firms cannot *all* raise their relative prices. Let the marginal cost of each firm i be given by $Pc(y_i)$, where y_i is the quantity supplied and P is the general price level. (Marginal costs are proportional to the general price level because the variable factors of production are supplied at *relative* prices that depend upon the quantity demanded of those factors.) Let us suppose furthermore that $c(y)$ is an increasing function, for the sort of reasons cited by Mitchell. Then an increase in the quantity supplied by industry i, if not associated with any shift in the marginal cost *schedule*, will be associated with an increase in marginal cost. In the case of an individual firm or industry, this need not be associated with any change in markups; it might simply be associated with an increase in the relative price P_i/P.[2]

But if we consider a uniform increase in the quantity produced by *each* sector, all relative prices P_i/P would have to increase by the same amount and this is not possible. In fact, in a symmetric equilibrium, one must have

$$\frac{1}{\mu} = c(Y) \tag{1.1}$$

where Y is the common (and hence aggregate) level of output, and μ the common (and hence average) markup. It follows from (1.1) that an increase in output Y is possible only insofar as either the real marginal cost schedule c shifts, or the markup falls. If firms allow their markups to decline, this will mean a higher level of equilibrium output than would otherwise be possible, given the current real marginal cost schedule. Thus if markups decline in the later stages of economic expansions, as Mitchell argues, this is not something that brings the expansion to an end; rather, it makes the expansion stronger (and possibly more prolonged) than is justified by cost conditions alone.[3]

Equation (1.1) suggests that a useful question about fluctuations in aggregate activity is to ask to what extent they result from variations in average markups, as opposed to shifts in the real marginal cost schedule of the typical firm. A related question that has received more attention in the literature is the extent to which the historical record actually suggests that markups were indeed low when output was high. This question is related to the first because, if the data suggested that markups were constant (or procyclical), independent movements in markups could not possibly account for a significant portion of output fluctuations. Thus, our survey of empirical evidence starts in section 2 with the simpler question of the extent to which measured markups are countercyclical. Since marginal cost cannot measured directly, all of these measures are indirect and rely on theories of the cost function facing individual firms. Once one has made these assumptions, however, one need only make a few additional suppositions to obtain an estimate of the derivative of the function c with respect to Y. This then allows one to infer the output fluctuations induced by changes in measured markups. the result is that one obtains a decomposition of output in terms of output movements due to markup changes and output movements due to shifts in the marginal cost schedule. We

consider this decomposition in section 3.

Section 4 is devoted to a brief survey of models of variable markups.[4] The above decomposition makes it clear that these models can serve two separate purposes. First, they can affect the extent to which shifts in the marginal cost schedule affect output. If, for example, reductions in marginal cost lower markups, their effect on output is magnified. What is perhaps more interesting still is that these theories allow shocks other than shocks to the marginal cost schedule to affect output as long as these shocks affect markups. In particular, allowing for endogenous markup variations adds a channel through which demand disturbances may affect output and employment.

This does not mean that the part of output variations that is due to shifts in the real MC schedule is due to "supply shocks", and the part due to markup variations with the part due to "demand shocks". There are various ways in which demand disturbances might, in principle, shift the real MC schedule.[5] Similarly, as we have already mentioned, the models of endogenous markup variation discussed below imply that "supply shocks" as well as "demand shocks" may cause markups to vary. For example, in Rotemberg and Woodford (1996b) it is shown how an endogenous increase in markups following an oil price increase can increase the contractionary impact of such a shock, even though the oil shock would also contract output (by shifting up the real MC schedule) to a lesser extent in a perfectly competitive model.

An objection to the usefulness of this decomposition might be that the validity of (1.1) does not itself imply that output variations are usefully "explained" by the markup variations that occur at the same time. One might view the markup as simply the ratio of two quantities (P and MC) that are each determined by other (and relatively independent) factors, including output, the causal determination of which must be understood in other ways. Such a view is possible if one views prices as evolving without reference to marginal costs (perhaps according a "Phillips curve" relation that makes the rate of price change a function of the level of real activity), and output as determined by demand given the current level of prices. In this case, markups might covary systematically with output (because of

(1.1)), but this would be irrelevant for output determination. Such a crude view, however, is difficult to take seriously. While we believe that nominal price stickiness plays a role, at least in short-run fluctuations in activity – and this is one of the reasons for markup variation taken up in section 4 below – it is not plausible that the level of marginal costs should not be a crucial determinant of the evolution of prices. If a firm's price is expected to remain fixed for a period of time, then the price chosen will depend not solely upon the current level of marginal cost, but upon (loosely speaking) the average of the expected levels of marginal cost over the entire period for which the price will be fixed. As a consequence, the rate of inflation (and expectations regarding future inflation) will be one of its determinants, as is explained further in section 4.1 below. In such a model, the connection between inflation and real activity (*i.e.*, the aggregate supply relation) can be usefully understood as resulting from the relation between inflation and the average markup, on the one hand, and the relationship between the markup and output determination indicated by (1.1).[6]

Furthermore, it does not seem that nominal rigidities alone can account for all markup variations. First, some of these variations appear not to be related to inflation in the way that can be accounted for by the simple hypothesis of prices remaining fixed for a time. For example, Rotemberg and Woodford (1996b) find that markups rise following an oil shock; but such shocks have also been associated with increases rather than decreases in inflation, so that slowness of prices to adjust to changes in nominal marginal costs might be expected to shrink markups, rather than raising them. And second, the very slowness of prices to adjust to changes in marginal costs (following, say, a loosening of monetary policy) is more easily explained if one posits that decreases in *desired* markups (*i.e.*, the ratio between price and marginal cost that would be chosen in the absence of the nominal price rigidity) coincide with declines in the ratio of actual to desired prices (due to slow price adjustment). If desired markups decline endogenously at times of temporarily high output, this "real rigidity" will amplify the effects of nominal rigidities, so that nominal disturbances have both larger and more persistent real effects. For both reasons, it would seem that variations in desired markups – and not simply variations in markups resulting from a discrepancy

between actual and desired prices – are of some importance in accounting for aggregate fluctuations. A consideration of the determinants of desired markups is therefore required.

Our survey proceeds as follows. In the next section, we discuss the evidence on cyclical variation in markups. Most of the evidence reviewed relates the slightly procyclical behavior of real wages to cyclical movements in the marginal product of labor. The key issue in subsections 2.1 and 2.2 is whether, as a result of technical progress, the marginal product of labor is as procyclical as real wages. While the marginal product of labor cannot be measured directly, we provide a number of reasons to believe that it is substantially more countercyclical than the relevant real wage so that, indeed, markups are countercyclical. In subsection 2.3 we consider measures of markup variation that are not based on wage variations; these involve cyclical variations in the use of intermediate inputs and in inventory accumulation. We then proceed to study responses of markups to particular shocks. Insofar as we are able to identify nontechnological disturbances, the analysis of markup changes is considerably simplified because we do not have to worry about the effect of technical progress on the real marginal cost schedule. Section 2 closes with an analysis of the differences in markup variations across industries.

Section 3 then turns to the consequences of markup variations for macroeconomic variables of interest. First we deal with the effect of markup changes on productivity and profits. We show that, under a variety of circumstances, increases in output that are caused by reductions in markups are associated with increases in profits and measured productivity. Since both productivity and profits are known to be procyclical, this is important in making sure that it is not implausible for changes in markups to be behind movements in output. Section 3 concludes with a method for decomposing output movements into those caused by shifts in the marginal cost schedule and movements due to markup changes (which induce movements along particular marginal cost curves). Section 4 is devoted to a survey of theories of markup variation, and section 5 concludes.

2 The Cyclical Behavior of Markups

In this section, we discuss empirical evidence regarding variation in markups over the business cycle. The main challenge in constructing measures of markup variation is to find suitable measures of marginal cost; and for this reason, it will often be useful to think, equivalently, of how one should measure cyclical variations in real marginal cost.[7] It is not easy to obtain measures of marginal cost of which one can be certain. Nonetheless, a variety of considerations may be offered, several of which provide support for the view that real marginal costs are procyclical, and hence that markups are countercyclical in the typical sector. If so, this implies that markup variations play a role in causing or at least amplifying cyclical fluctuations in economic activity.

2.1 Cyclical Behavior of the Labor Share

The most common measures of marginal cost in the literature consider the cost of increasing output through an increase in the labor input, holding fixed other inputs. (Measures of marginal cost deriving from variation of production decisions along other margins are considered in section 2.7 below. Note that if firms are minimizing cost, the measure of marginal cost obtained by considering each possible margin should be the same, so that it suffices to consider one.) If output is a differentiable function of the labor input, and firms are wage-takers, then marginal cost is equal to the wage divided by the marginal product of labor. If we assume an aggregate production function of the form

$$Y = F(K, zH), \tag{2.1}$$

where K is the capital stock, H the number of hours worked, and z an index of labor-augmenting technical progress, then the markup of price over marginal cost μ is given by[8]

$$\mu = \frac{PzF_H(K, zH)}{W}. \tag{2.2}$$

This equation provides an approach to the measurement of markup (or real marginal cost) variations. It also highlights two reasons for the real marginal cost schedule referred to

in section 1 to be upward-sloping. The first is that, holding constant other determinants of labor supply, the real wage must presumably rise to induce more people to work. The second is that, if one makes the standard assumption that the production function F is concave and one fixes both the capital stock and the state of technology z, the marginal product F_H is a decreasing function of the labor input.

Whether or not typical increases in employment are in fact associated with markup declines depends, however, upon whether they are associated with increases in K or z, or decreases in the real wage W/P, sufficient to offset the effects of the increase in the labor input on F_H. In general, real wages do not move countercyclically, and in fact, there is clearly *procyclical* variation in the real wages received by individuals, once one corrects for cyclical variation in the composition of the workforce.[9] This is the famous criticism raised by Dunlop (1938) and Tarshis (1939) against the theory of aggregate supply of Keynes (1936), which is essentially just (2.2), under the assumption of constant technology and a markup equal to one. Keynes (1939) recognized the appeal of a hypothesis of countercyclical markup variation as a resolution of the puzzle, which is the interpretation that we propose.

An alternative explanation, of course, is that real wages are procyclical because fluctuations in activity are caused by variations in technical progress. (Real business cycle models are sometimes criticized for predicting real wage movements that are *too* procyclical, so that Kydland and Prescott, 1988, take their findings as support for the technology-shock hypothesis.) The need to correct for possible variations in the rate of technical progress means that one needs to measure the variation in both the labor input and in the quantity produced. The required calculation is especially simple if, following Bils (1987), we specialize the production function (2.1) to the case

$$Y = g(K)(zH)^{\alpha}, \tag{2.3}$$

where g is a positive increasing function, and $\alpha > 0$. Marginal cost is then $WH/\alpha Y$, so that the markup is given by

$$\mu = \alpha s_H^{-1}, \tag{2.4}$$

where s_H is the labor share WH/PY. Under these assumptions, markup variations are simply the inverse of the variations observed in the labor share. In the case of a Cobb-Douglas production function (or the slightly more general form assumed above), marginal cost is proportional to average labor cost so that a valid measure of markup variations is given by fluctuations in the ratio of price to "unit labor cost" WH/Y – a measure of variations in price-cost margins often referred to in empirical studies of business cycles such as those of Moore (1983).[10]

We first consider the evidence regarding cyclical variation in this simple measure. The price P with which firms are concerned is the price they receive for their products. This means that the relevant labor share is not the ratio of labor compensation to the value of output conventionally measured in national income accounts, but rather the ratio to the revenue received by firms, which equals the value of output minus indirect taxes.[11] We consider cyclical variation in three different measures of this labor share, for the whole economy, the corporate sector and the nonfinancial corporate sector respectively. The first of these measures is less satisfactory than the others for two reasons. First, it includes the government, many of whose services are not sold in markets. Second, it includes income of proprietors in the denominator, and this contains an element of compensation as well. The use of the narrower measures of the labor share eliminates both of these problems. Nonetheless, we include some statistics relating to the overall labor share as well, for comparability with other studies.

Figure 1 plots these three series for the period 1947:1 to 1993:1.[12] For future reference, the figure also plots the Hodrick-Prescott trend in the labor share for the nonfinancial corporate sector. The figure reveals that the labor share in the corporate sector is essentially identical to the labor share in the nonfinancial segment. On the other hand, the overall labor share deviates from the other two shares in the early 1960's and remains above them from then on. All three series show large increases in the late 1960's. Particularly for the labor share corresponding to the whole economy, this appears to represent a structural break that cannot be regarded as an example of business-cycle variation in the series. Hence in considering the

cyclical behavior of the series, we also considered a sample that begins only in 1970.

We next wish to relate the movements in one or another of these labor shares to those of a variable that measures the business cycle. One attempt to do so is provided in Figure 2, which plots the labor share in the nonfinancial corporate business sector against the NBER recessions. For each of these recessions, the first line in Figure 2 represents a business cycle peak while the second represents the trough. At first glance, this picture might suggest that the labor share is countercyclical because the labor share series tends to have a local maximum between peaks and troughs. But, it is important to remember that, for the labor share to be perfectly procyclical, its peaks ought to be aligned with the business cycle peaks themselves. The labor share ought then to decline between peak and trough and then increase during the recovery that takes place after the trough. It is this latest implication of a procyclical labor share that is closest to being true, as the labor share increases in the recoveries after the 1954, 1974 and 1982 recessions.

The actual correlation between the change in the labor share and a dummy variable that takes the value of one between peaks and troughs and a value of zero otherwise is .17, which suggests, albeit very weakly, that the labor share rises when output is declining. However, the correlation between the labor share and the two quarter lagged value of this dummy variable is -.28. Thus, as the plot itself suggests, the labor share tends to rise late in expansions and to fall late in recessions. This basic pattern: a weak slightly negative relation between the labor share and contemporaneous cyclical indicators and a much stronger positive relation between the labor share and slightly lagged indicators of cyclical activity is what comes out of a more formal analysis as well.

For this slightly more formal statistical analysis, we considered four indicators of the business cycle. A popular indicator of this sort is obtained by detrending real GDP using the Hodrick-Prescott filter and then using this detrended series to be a measure of the business cycle. This is, however, a rather arbitrary procedure (since there is no obvious reason to choose one value rather than another of the weighting parameter that determines the degree to which the trend is smoothed).[13] An alternative that we find appealing is to

follow Beveridge and Nelson (1980) and equate the "cyclical" level of GDP at time t with the expected decline in GDP from time t onwards. This captures the intuitive idea that cyclical movements are temporary, so that a cyclically low level of output corresponds to a high expected rate of growth of output. The difficulty with this approach is that it only becomes meaningful when one specifies an information set that can be used to forecast GDP growth. Following on the steps of an extensive literature, Rotemberg and Woodford (1996a) show that the linearly detrended level of hours spent working in nonagricultural establishments and the ratio of consumer expenditures on nondurables and services to GDP are particularly useful in this respect.

A simpler cyclical indicator is the linearly detrended level of hours worked. We prefer to linearly detrend hours rather than GDP, since once a trend is included in the regression, the Dickey-Fuller test strongly rejects the hypothesis that the logarithm of hours worked in non-agricultural establishments has a unit root. This measure of detrended hours is in fact one of the main components of the Rotemberg-Woodford measure of forecastable output movements; low levels of hours worked (which are closely related to high unemployment), imply that output can be expected to grow and are thus a good indicator of recessions. For purposes of comparison, we also consider an hours series that has been detrended using the Hodrick-Prescott filter.

The first three rows of Table 1 report correlations of our three measures of the labor share with the various cyclical indicators, over the sample period 1947:1 through 1993:1. The first column shows correlations with the predicted declines in output over 12 quarters considered in Rotemberg and Woodford (1996a), the second column shows the correlations with the Hodrick-Prescott filtered level of output, the third column uses linearly detrended hours while the last uses the hours series detrended using the H-P filter. Except the correlations with linearly detrended hours (which are small and positive), the other correlations are small and negative; suggesting weak countercyclical movements in the labor share.

These results are consistent with Boldrin and Horvath (1995), Gomme and Greenwood (1995) and Ambler and Cardia (1996) who also report negative correlations of the labor share

with output. The correlations they report are larger in absolute magnitude because they use the H-P filter to detrend the labor share as well as to detrend output. Using the H-P filter to detrend the labor share seems problematic, however, because the large movements at relatively high frequencies of the resulting "trend" displayed in Figure 1 are difficult to interpret.[14]

The last twelve rows of Table 1 report the correlations of the labor share in the nonfinancial corporate sector with leads and lags of the four cyclical indicators. The correlations with the lags are uniformly positive and attain their biggest value when the cyclical indicator is lagged four quarters. Thus a high level of activity is associated with subsequent increases in the labor share. Interestingly, this result also extends to the case where we study the cross-correlogram of H-P filtered output and the H-P filtered labor share. The correlations of the labor share with the leads of our cyclical indicators are uniformly negative. This means that the labor share peaks before the peak in hours.

2.2 Corrections to the Labor-Share Measure of Real Marginal Cost

While the labor share (or equivalently, the ratio of price to unit labor cost) is a familiar and easily interpretable statistic, it represents a valid measure of markup variations only under relatively special assumptions. In this section, we briefly discuss a number of corrections to this measure that would arguably be required to obtain a more realistic measure of real marginal cost. As we shall see, several of these corrections imply that real marginal cost is more procyclical than the labor share. For any of several reasons, then, the measurements discussed above may understate the degree to which cyclical variations in output and employment are due to markup variations as opposed to shifts in the real marginal cost schedule. We take the possible corrections up in sequence.

A Non-Cobb-Douglas Production Function.

Suppose that the aggregate production function is of the general form (2.1), but that F is not necessarily of the Cobb-Douglas form (or, more precisely, isoelastic in the labor input,

as in (2.3)). Equation (2.2) can still equivalently be written

$$\mu = \eta_H s_H^{-1}, \tag{2.5}$$

where $\eta_H \equiv zHF_H(K, zH)/F(K, zH)$ is the elasticity of output with respect to the (effective) labor input. (Equation (2.5) reduces to (2.4) in the case of a constant elasticity.)

The effect of the additional variable factor in (2.5) depends upon the nature of cyclical variations, if any, in the elasticity of output with respect to the labor input. In the case that F exhibits constant returns to scale, the elasticity η_H can be expressed as a function of the effective labor-to-capital ratio, zH/K, or equivalently as a function of the output-to-capital ratio:

$$\eta_H = \eta_H(y), \tag{2.6}$$

where $y \equiv Y/K$. In the case that the elasticity of substitution between capital and (effective) labor inputs is less than one, the function $\eta_H(y)$ is monotonically decreasing. This would seem the most likely direction of deviation from the Cobb-Douglas case (a constant elasticity of substitution exactly equal to one), as the Cobb-Douglas specification is widely regarded as a reasonable representation of long-run substitution opportunities, whereas short-run factor substitutability (which is relevant for the present calculation) might well be less (for example, because technology is "putty-clay"). In this case, because y is a procyclical variable (this follows directly from the fact that the capital stock evolves slowly relative to the length of business-cycle fluctuations), (2.6) implies that the additional factor η_H in (2.5) imparts additional countercyclical variation to the implied markup series, roughly coincident with the cyclical component of output or hours. A correction of this kind thus leads to the conclusion that markups fall more in booms than is suggested by the simple labor share measure, and that markup declines coincide more closely in time with increases in output and hours.

The size of this correction can be quantified as follows. Assuming constant returns to scale, the elasticity of η_H with respect to y is given by $a \equiv (1 - \epsilon_{KH}^{-1})(\eta_H^{-1} - 1)$, where ϵ_{KH} is the (Hicks-Allen) elasticity of substitution between capital and labor inputs. A log-linear

approximation to the markup series implied by (2.5) is then given by

$$\hat{\mu} = a\hat{y} - \hat{s_H}, \tag{2.7}$$

where hats denote deviations of the logarithm of a stationary variable from its average value. A quantitative estimate of the elasticity a requires values for ϵ_{KH} and for the average value of η_H (*i.e.*, the value of this elasticity in the case of the "steady-state" factor ratio around which one considers perturbations). Using (2.5), the latter parameter may be calibrated from the average labor share, given an estimate of the average markup. resulting in $a = (1-\epsilon_{KH}^{-1})(\mu^{-1}s_H^{-1}-1)$. (In this last expression, all symbols refer to the average or steady-state values of the variables.) With a markup μ near one, a labor share of .7 and an elasticity of substitution ϵ_{KH} of .5, this formula gives a value of a equal to -.4. Table 2 reports the resulting correlations between $\hat{\mu}$ and predicted declines in GDP for markups based on both the nonfinancial corporate and the private labor shares. Not surprisingly, markups are now much more countercyclical. However, the contemporaneous correlation of the markup with detrended output is still smaller in absolute value than the correlation with lagged output. Also, as in the case where we do not adjust the labor share, the correlations with leads of output are greater than the contemporaneous correlation. These are actually positive for output led more than 3 quarters.[15]

Overhead Labor.

In deriving (2.5) – (2.7), we assume constant returns to scale. An important reason why this may be inaccurate is the presence of overhead costs. Particularly relevant to the above calculations would be the existence of overhead labor. Suppose that each firm's production function[16] is of the form $Y = F(K, z(H - \bar{H}))$, where F is homogeneous of degree one as before, and $\bar{H} \geq 0$ represents "overhead labor" that must be hired regardless of the quantity of output that is produced. Note that an overhead labor requirement implies increasing returns (average cost exceeding marginal cost), although marginal cost remains independent of scale.[17]

Replacing (2.1) by this, implies that (2.5) should become instead

$$\mu = \Big(\frac{H}{H - \bar{H}}\Big)\eta_H s_H^{-1}, \tag{2.8}$$

where η_H now refers to the elasticity of output with respect to the effective *non-overhead* labor input. Under this definition, η_H is again constant in the case that F takes the Cobb-Douglas form, and countercyclical if the elasticity of substitution between capital and labor is less than one. The new factor in (2.8), $\frac{H}{H-\bar{H}}$, is a monotonically decreasing function of H if $\bar{H} > 0$.[18] Allowing for overhead labor thus provides a further reason to regard markups as more countercyclical than is indicated by the labor share alone. A similar conclusion is reached if one assumes fixed costs in production that do not take the form of overhead labor alone, *e.g.*, if one assumes a production function of the form

$$Y = F(K, zH) - \Phi,$$

where $\Phi > 0$ represents the fixed costs of operation.

The consequences of this correction can be quantified as follows. The elasticity of the factor $\frac{H}{H-\bar{H}}$ with respect to H is given by $b \equiv -s_o/(1 - s_o)$, where s_o is the average or steady-state value of $\bar{H}/H$, the share of overhead labor in the total labor input. Equation (2.7) may then be generalized to yield

$$\hat{\mu} = a\hat{y} + b\hat{H} - \hat{s_H}. \tag{2.9}$$

The elasticity b is obviously non-positive. Its size depends on the average fraction of labor which constitutes overhead labor. A useful bound on this can be obtained by relating s_o to the degree of returns to scale. Let the index of returns to scale ρ be defined as the ratio of average cost to marginal cost of production. Measured at the steady-state factor inputs, one obtains

$$\rho = 1 + \eta_H\Big(\frac{s_o}{1 - s_o}\Big),$$

so that instead of calibrating s_o, one may equivalently calibrate ρ. In terms of this parameter, we obtain $\eta_H = \mu s_H - (\rho - 1)$ for the steady-state elasticity of output with respect to non-overhead labor, and $b = -(\rho - 1)/[\mu s_H - (\rho - 1)]$.

It is easily seen that one must have $\rho \leq \mu$, in order for there to exist non-negative profits in the steady state. This allows one to bound the possible size of the elasticity b, given an estimate of the average markup. On the other hand, the same consideration provides a reason for supposing that overhead costs are non-negligible, if one believes that prices do exceed marginal cost on average. For with constant returns to scale, prices higher than marginal cost would imply the existence of pure profits (in addition to the competitive return to capital); for example, a markup of 25% ($\mu = 1.25$) would imply that pure profits should make up 20% of total revenues. This is rather large given the scant evidence for the existence of pure profits in U.S. industry. Indeed, Hall (1988) finds (using stock market returns to construct a user cost for capital) that pure profits in U.S. industry are close to zero. It furthermore makes sense that profits should be zero in the steady state, due to entry, which one should expect to eliminate persistent profits in the long run, even if entry does not respond quickly enough to eliminate cyclical fluctuations in profits. If we assume this, we can impose $\rho = \mu$, so that there is only a single parameter to calibrate, that describes both the degree of returns to scale and the degree of market power. With $\mu = 1.25$ and a labor share of .7, the parameters b is then -.4. Table 2 shows that, even letting a equal zero, such a value of b leads to markups that are strongly countercyclical though the correlations with lagged output remain higher in absolute value.[19]

The significance that one attaches to such findings obviously depends upon the size of the average markup (or degree of returns to scale) that one is willing to assume. Here it is worth remarking that a value of μ equal to 1.6 need not mean that any individual firm marks up its costs by 60%. The reason for this is that firms do not just mark up their labor costs but also their materials cost. To see what this implies about the markup, suppose that, as in Rotemberg and Woodford (1995), materials are a fixed proportion s_M of aggregate output while value added constitutes only a fraction $(1 - s_M)$ of total costs. The marginal cost of producing one unit of gross output is then

$$\frac{(1 - s_M)W}{zF_H} + s_M$$

and the markup of the price of gross output over total marginal cost μ^{GO} is given by

$$\frac{1}{\mu^{GO}} = (1 - s_M)\frac{1}{\mu^{VA}} + s_M, \tag{2.10}$$

where μ^{VA} is the "value-added markup" that satisfies (2.2). If the materials share equals 0.6 (as is typical of U.S. manufacturing), then a μ^{VA} of 1.6 (the "baseline case" of Rotemberg and Woodford, 1991) requires that the typical *firm's* price be only 18% higher than its marginal cost.

A related correction would assume, instead of overhead labor, a "setup cost" for each employee, as is considered in Basu and Kimball (1994). Suppose that the production function is $Y = F(K, z(h-\bar{h})N))$, where now N represents the number of employees and h the number of hours worked by each. We again assume that F is homogeneous of degree one; the "set-up cost" $\bar{h} > 0$ represents a sort of per-employee fixed cost. (The observed preference for full-time employees observed in many lines of work makes the existence of such costs plausible.[20]) If we consider the marginal cost of increasing output solely on the employment margin (holding fixed hours per week), we again obtain (2.8), but with H and $\bar{H}$ replaced by h and $\bar{h}$ in the first factor. We correspondingly again obtain (2.9), but with $\hat{H}$ replaced by $\hat{h}$. Since hours per employee are also a strongly procyclical variable, the first factor in (2.8) is again a source of further countercyclical movement in implied markups. Basu and Kimball suggest that $s_o = .25$ should be an upper bound on the importance of such set-up costs (as full-time wage premia should otherwise be larger); but this value would still allow the elasticity in (2.9) to be as large as $b = -0.3$.

Marginal Wage Not Equal to the Average.

Thus far, we have assumed wage-taking behavior on the part of firms, meaning that they regard themselves as being able to hire additional hours of work, at the margin, at a wage which is also the wage paid for each of the hours that they do hire – so that the relevant marginal wage is also the average wage that is paid. Suppose, however, that this is not true, and that the firm's wage bill is $W(H)$, a function that is increasing, but not necessarily linear in H.[21] In this case, marginal cost depends upon the *marginal* wage, $W'(H)$, so that

(2.5) becomes

$$\mu = \omega^{-1}\eta_H s_H^{-1}. \tag{2.11}$$

where $\omega \equiv HW'(H)/W(H)$ is the ratio of the marginal wage to the average wage. This might vary cyclically for several reasons.

One reason might be monopsony power in the labor market. Suppose that each firm faces an upward-sloping firm-specific labor supply curve, and takes this into account in its hiring and production decisions. (The wage that the firm must pay may also depend upon other variables such as the overall level of employment in the economy, but these factors are taken as given by the individual firm, and can simply be treated as time-variation in the location of the firm-specific labor supply curve.) If $w(H)$ is the wage that the firm must pay if it hires H hours of work, then $W(H) = Hw(H)$, and $\omega = 1 + \epsilon_{Hw}^{-1}$, where ϵ_{Hw} is the elasticity of the firm-specific labor supply curve. This might be either increasing or decreasing with increases in hours hired by the firm. The most plausible assumption, however, would probably be that the elasticity of labor supply decreases as the hours hired by the firm increase (it is hard to induce people to work more than a certain number of hours, even at very high wages, while on the other hand the opportunity cost of their time tends not to fall below a certain level even when the number of hours worked is small). Under this assumption, the factor ω is an increasing function of H, and (2.9) again holds, with $b < 0$. This would imply that real marginal costs would actually be more procyclical (and markups more countercyclical) than would be suggested by consideration only of the terms in (2.5).

Alternatively, one might imagine that firms first hire a certain number of employees, and that they initially contract with them about a wage *schedule* which determines the wage as a function of hours worked. Subsequently, perhaps after receiving additional information about current demand conditions, the firms determine the hours of work. If all employees are asked to work the same number of hours at this stage, we may interpret $W(H)$ in (2.11) as the wage schedule negotiated with each employee. Now if the number of employees is chosen *ex ante* so as to minimize the cost of the number of hours that the firm expects to use, then *ex ante* expected hours per worker will be the level H^* that minimizes the average

wage $W(H)/H$.[22] At this point, the marginal wage should equal the average wage, and (assuming a unique minimum) in the case of small fluctuations in H around the value H^*, ω should be increasing in H. Again this would imply markups more countercyclical than would be suggested by (2.5).

Most observed wage contracts do not involve wages that increase continuously with the number of hours that the employee is asked to work. On the other hand, if one supposes that contractual wages are not the true shadow price of additional labor to a firm, because of the presence of implicit contracts of the kind assumed, for example, by Hall (1980), then one might suppose that the true cost to the firm rises in proportion to the employee's disutility of working, even if the wages that are paid in the current period do not. This would be a reason to expect the effective wage schedule $W(H)$ to be convex, so that the above analysis would apply.

Bils (1987) observes that in many industries, a higher wage is paid for overtime hours (*i.e.*, , hours in excess of 40 hours per week). He thus proposes to quantify the extent to which the marginal wage rises as firms ask their employees to work longer hours, by measuring the extent to which the average number of overtime hours per employee, V, rises with increases in the total number of hours worked per employee H, and then assuming that $W(H) = w_0[H + pV(H)]$, where w_0 is the straight-time wage and p is the overtime premium (0.5 according to the U.S. statutory requirement).[23] For example, he finds that when average hours per employee rise from 40 hours per week to 41 hours, the average number of overtime hours worked per employee rises by nearly 0.4 hours, while when they rise from 41 to 42 hours per week, overtime hours rise by another 0.5 hours. This increase in the fraction of hours that are overtime hours as average hours increase means not only that the marginal wage exceeds the average wage, but that the ratio of the marginal wage to the average wage rises as hours increase. Assuming $p = .5$, Bils finds that an increase in average hours from 40 to 41 hours increases the average wage by about 0.5%, but increases the marginal wage by 4.6%. On average, he finds that the factor ω in (2.11) has an elasticity of 1.4 with respect to variations in average hours.[24] Thus a log-linear approximation to (2.11) is again of the

form (2.9), where in Bils' work $\hat{H}$ refers to fluctuations in average hours per worker,[25] and $b = -1.4$.

Since average hours worked in U.S. manufacturing are strongly procyclical, taking into account this factor makes the implied markup significantly more countercyclical. Indeed, when Bils regresses his constructed markup series (using (2.9)) on a measure of cyclical employment,[26] he finds that markups decline, on average, by 0.33% for each one-percent increase in employment. Of this cyclical variation, a 0.12% decline is implied by the increase in the labor share (which is mildly procyclical in his sample), while the remaining 0.21% decline comes from the increase in the ratio of the marginal wage to the average wage.

One may question whether the statutory premium paid for overtime hours represents a true cost to the firm; some argue, for example, that the opportunity to work overtime is in fact dispensed as a reward for exemplary behavior at other times. Bils answers this objection by pointing out that if one assumes that because of sophisticated implicit contracts, the true cost to the firm is proportional to the worker's disutility of working $v(H)$, then one might well obtain estimates of the degree of procyclical movement in the ratio of the marginal wage to the average that are as large as those obtained using his method. Under the assumption suggested above about the steady-state level of hours, the coefficient b in (2.9) would in that case equal $-v''/H^*v'$, or $-\epsilon_{Hw}$, where ϵ_{Hw} is now the Frisch (or intertemporal) elasticity of labor supply by a wage-taking household in a competitive spot market. A value of b less negative than Bils' value of -1.4 would then be obtained only if one assumed preferences implying an elasticity of labor supply greater than 0.7, whereas many microeconomic studies of labor supply estimate a lower elasticity than that.

Costs of Adjusting the Labor Input.

An additional reason why marginal hours may be more expensive in booms is the presence of adjustment costs. It is simplest to illustrate this point if we assume, as, for example, in Pindyck and Rotemberg (1983), that there are convex costs of changing the labor input H. Suppose that, in addition to the direct wage costs $w_t H_t$ of hiring H_t hours in period t,

there is an adjustment cost of $\kappa_t H_t \phi(H_t/H_{t-1})$. Here κ_t represents a price index in period t for the inputs that must be purchased as part of the adjustment process; we shall assume that the (logarithms of the) factor prices κ and w are co-integrated, even if each is only difference-stationary. (More specifically, we shall assume that κ/w is stationary.) The factor $H_t\phi(H_t/H_{t-1})$ represents the physical quantity of inputs that must be expended in order to adjust the labor input; note that adjustment costs increase in proportion to the quantity of labor used by a given firm. This specification implies that adjustment costs remain of the same magnitude relative to direct labor costs, even if both H and w exhibit (deterministic or stochastic) trend growth. The exposition is simplest if we treat the adjustment costs as "external", in the sense that the additional inputs that must be purchased are something other than additional labor, so that both the production function (2.1) and the formula for the labor share can still be written as before in terms of a single state variable "H".[27] Finally, we assume that ϕ is a convex function, with $\phi(1) = \phi'(1) = 0$; thus adjustment costs are non-negative, and minimized (equal to zero) in the case of no change in the labor input.

We can then compute the marginal cost associated with an increase in output at date t, assuming that production is increased solely through an increase in the labor input at date t, with no change in the inputs used in production at other dates, except for the necessary changes in the inputs used in the adjustment process at both dates t and $t+1$. In this case, (2.5) becomes[28]

$$\mu = \Omega^{-1}\eta_H s_H^{-1}, \tag{2.12}$$

where

$$\Omega_t = 1 + (\kappa_t/w_t)\{[\phi(\gamma_{Ht}) + \gamma_{Ht}\phi'(\gamma_{Ht})] - E_t[R_{t,t+1}\gamma_{\kappa t+1}\gamma_{Ht+1}^2\phi'(\gamma_{Ht+1})]\}, \tag{2.13}$$

in which in turn $\gamma_{Ht} \equiv H_t/H_{t-1}$, $\gamma_{\kappa t} \equiv \kappa_t/\kappa_{t-1}$,[29] and $R_{t,t+1}$ is the stochastic discount factor by which firms discount random income at date $t+1$ back to date t. (Here we have written (2.13) solely in terms of variables that we expect to be stationary, even if there are unit roots in both H and w, to indicate that we expect Ω to be a stationary random variable.

If ϕ is strictly convex (*i.e.*, if there are non-zero adjustment costs), the cyclical variation

in the factor Ω changes the nature of implied markup fluctuations. Because ϕ' is positive when the labor input is rising and negative when it is falling, Ω should be a procyclical factor, though with a less exact coincidence with standard business cycle indicators than the cyclical correction factors discussed thus far. If we take a log-linear approximation to (2.13), near a steady-state in which the variables $H, \kappa/w, \gamma_\kappa$, and R are constant over time, we obtain

$$\hat{\Omega}_t = c[\hat{\gamma}_{Ht} - \beta E_t \hat{\gamma}_{Ht+1}], \tag{2.14}$$

where here the coefficient $c > 0$ denotes $\phi''(1)$ times the steady state value of κ/w, and β denotes the steady-state value of $R\gamma_\kappa$, the discount factor for income streams measured in units of the adjustment-cost input. This can then be substituted into the log-linear approximation to (2.12),

$$\hat{\mu} = a\hat{y} - \hat{s_H} - \hat{\Omega}, \tag{2.15}$$

to obtain a formula to be used in computing markup variations. Equation (2.14) makes it clear that the cyclical variations in the labor input are the main determinant of the cyclical variations in Ω. The factor Ω will tend to be high when hours are temporarily high (both because they have risen relative to the past and because they are expected to fall in the future), and correspondingly low when they are temporarily low. Thus, it tends to increase the degree to which implied markups are countercyclical.[30]

More precisely, the factor Ω tends to introduce a greater negative correlation between measured markups and *future* hours. Consider, as a simple example, the case in which hours follow a stationary AR(1) process given by

$$\hat{H}_t = \rho \hat{H}_{t-1} + \epsilon_t,$$

where $0 < \rho < 1$, and ϵ is a white-noise process. Then $\hat{\Omega}_t$ is a positive multiple of $\hat{H}_t - \lambda \hat{H}_{t-1}$, where $\lambda \equiv (1 - \beta(1-\rho))^{-1}$, and $\text{cov}(\hat{\Omega}_t, \hat{H}_{t+j})$ is of the form $C(1-\lambda\rho)\rho^j$ for all $j \geq 0$, where $C > 0$, while it is of the form $C(1-\lambda\rho^{-1})\rho^{-j}$ for all $j < 0$. One observes (since $\rho < \lambda < 1/\rho$) that the correlation is positive for all leads $j \geq 0$, but negative for all lags $j < 0$. Thus this correction would make the implied markup series more negatively correlated with leads of

hours, but less negatively correlated with lags of hours. The intuition for this result is that high lagged levels of hours imply that the current cost of producing an additional unit is relatively low (because adjustment costs are low) so that current markups must be relatively high. Since, as we showed earlier, the labor share is more positively correlated with lags of hours (and more negatively correlated with leads of hours) this correction tends to make computed markup fluctuations more nearly coincident with fluctuations in hours. To put this differently, consider the peak of the business cycle where hours are still rising but expected future hours are low. This correction suggests that marginal cost are particular high at this time because there is little future benefit from the hours that are currently being added.

The last two columns of Table 2 show the effect of this correction for c equal to 4 and 8 while β is equal to .99. To carry out this analysis, we need an estimate of $E_t\hat{\gamma}_{Ht+1}$. We obtained this estimate by using one of the regressions used to compute expected output growth in Rotemberg and Woodford (1996a). In particular, the expectation at t of $\hat{H}_{t+1}$ is the fitted value of a regression of $\hat{H}_{t+1}$ on the values at t and $t-1$ of $\hat{H}$, the rate of growth of private value added and the ratio of consumption of nondurables and services to GDP. Subtracting the actual value of $\hat{H}_t$ from this fitted value, we obtain $E_t\hat{\gamma}_{Ht+1}$. This correction makes the markup strongly countercyclical and ensures that the correlation of the markup with the contemporaneous value of the cyclical indicator is larger in absolute value than the correlation with lagged values of this indicator. On the other hand, the correlation with leads of the indicator is both negative and larger still in absolute value, particularly when c is equal to 8.

The same calculations apply, to a log-linear approximation, in the case that the adjustment costs take the form of less output from a given quantity of labor inputs. Suppose that in the above description of production costs, H refers to the hours that are used for production purposes in a given period, while $H\phi$ indicates the number of hours that employees must work on tasks that are created by a firm's variation of its labor input over time. (In this case, $\kappa \equiv w$.) Equations (2.12) – (2.13) continue to apply, as long as one recalls that H and s_H now refer solely to hours used directly in production. Total hours worked equal AH

instead, and the total labor share equals As_H, where $A \equiv 1 + \phi(\gamma_H)$. But in the log-linear approximation, we obtain $\hat{A} = 0$, and so equations (2.14) – (2.15) still apply, even if $\hat{\gamma}_H$ and $\hat{s_H}$ refer to fluctuations in the *total* labor inputs hired by firms.

A more realistic specification of adjustment costs would assume costs of adjusting *employment*, rather than costs of adjusting the total labor input as above.[31] Indeed, theoretical discussions that assume convex costs of adjusting the labor input, as above, generally motivate such a model by assuming that the hours worked per employee cannot be varied, so that the adjustment costs are in fact costs of varying employment. In the data, however, employment variations and variations in total person-hours are not the same, even if they are highly correlated at business-cycle frequencies. This leads us to suppose that firms can vary both employment N and hours per employee h, with output given by $F(K, zhN)$, and that costs of adjusting employment in period t are given by $\kappa_t N_t \phi(N_t/N_{t-1})$. If, however, there are no costs of adjusting hours, and wage costs are linear in the number of person-hours hired Nh, firms will have no need ever to change their number of employees (which is clearly not the case). If, then, one is not to assume costs of adjusting hours per employee,[32] one needs to assume some other motive for smoothing hours per employee, such as the sort of non-linear wage schedule discussed above. We thus assume that a firm's wage costs are equal to $W(h)N$, where $W(h)$ is an increasing, convex function as above.

One can then again compute the marginal cost of increased output at some date, assuming that it is achieved through an increase in employment at that date only, holding fixed the number of hours per employee h at all dates, as well as other inputs. One again obtains (2.12), except that the definition of Ω in (2.13) must be modified to replace γ_H by γ_N, the growth rate of employment, throughout. (In the modified (2.13), w now refers to the average wage, $W(h)/h$.) Correspondingly, (2.15) is unchanged, while (2.14) becomes

$$\hat{\Omega}_t = c[\hat{\gamma}_{Nt} - \beta E_t \hat{\gamma}_{Nt+1}], \tag{2.16}$$

Thus one obtains, as in the simpler case above, a correction to (2.5) that results in the implied markup series being more countercyclical (since employment is strongly procyclical,

just as with the total labor input).

Alternatively, one could compute the marginal cost of increased output, assuming that it is achieved solely through an increase in hours per employee, with no change in employment or in other inputs. In this case, one obtains again (2.11), but with H everywhere replaced by h in the first factor on the right-hand side. There is no contradiction between these two conclusions. For the right-hand sides of (2.11) and (2.12) should be equal at all times; cost-minimization requires that

$$W'(h_t) = w_t + \kappa_t[\phi(\gamma_{Nt}) + \gamma_{Nt}\phi'(\gamma_{Nt})] - E_t[R_{t,t+1}\kappa_{t+1}\gamma_{Nt+1}^2\phi'(\gamma_{Nt+1})]\}, \tag{2.17}$$

which implies that $\Omega = \omega$. Condition (2.17) is in fact the Euler equation that Bils (1987) estimates in his "second method" of determining the cyclicality of the marginal wage; he uses data on employment and hours variations to estimate the parameters of this equation, including the parameters of the wage schedule $W(h)$.[33] An equivalent method for determining the cyclicality of markups would thus be to determine the importance of employment adjustment costs from estimation of (2.17), and compute the implied markup variations using (2.15) – (2.16). Insofar as the specification (2.17) is consistent with the data, both approaches should yield the same implied markup series. It follows that Bils' results using his second method give an indication of the size of the correction that would result from taking account of adjustment costs for employment, if these are of the size that he estimated. His estimate of these adjustment costs imply an elasticity of Ω even greater than the value of 1.4 discussed above.

Labor Hoarding.

Suppose now that not all employees on a firm's payroll are used to produce current output at each point in time. For example, suppose that of the H hours paid for by the firm at a given time, H_m of these are used in some other way (let us say, maintenance of the firm's capital), while the remaining $H - H_m$ are used to produce the firm's product. Output is then given by $Y = F(K, z(H - H_m))$ rather than (2.1). We can again compute the marginal cost of increasing output by hiring additional hours, holding H_m fixed (along with other inputs).

One then obtains instead of (2.5)

$$\mu = u_H^{-1} \eta_H s_H^{-1}, \tag{2.18}$$

where $u_H \equiv (H - H_m)/H$ is the fraction of the labor input that is utilized in production. Note that this conclusion is quite independent of how we specify the value to the firm of the alternative use to which the hours H_m may be put. It suffices that we believe that the firm is profit-maximizing, in its decision to allocate the hours that it purchases in this way, as in its other input decisions, so that the marginal cost of increasing production by shifting labor inputs away from maintenance work is the same as the cost of increasing production by hiring additional labor.

The fraction u_H is often argued to be procyclical, insofar as firms are said to "hoard labor" during downturns in production, failing to reduce payrolls to the extent of the decline in the labor needed to produce their output, so as not to have to increase employment by as much as the firms' labor needs increase when output increases again. For example, the survey by Fay and Medoff (1985) finds that when output falls by one percent, labor hours used in production actually fall by 1.17 percent, but hours paid for fall only by 0.82 percent.[34]

Insofar as this is true, it provides a further reason why markups are more countercyclical than would be indicated by (2.5) alone.[35] If the Fay and Medoff numbers are correct, and we assume furthermore that nearly all hours paid for are used in production except during business downturns, they suggest that u_H falls when output falls, with an elasticity of 0.35 (or an elasticity of about 0.4 with respect to declines in reported hours). Thus this factor alone would justify setting $b = -0.4$ in (2.9).

A related idea is the hypothesis that effective labor inputs vary procyclically more than do reported hours because of procyclical variation in work effort. We may suppose in this case that output is given by $Y = F(K, zeH)$, where e denotes the level of effort exerted. If, however, the cost of a marginal hour (which would represent e units of effective labor) is given by the reported hourly wage W, then equation (2.5) continues to apply. Here the presence of time-variation in the factor e has effects that are no different than those of time-variation in the factor z, both of which represent changes in the productivity of hours worked; the fact

that e may be a choice variable of the firm while z is not has no effect upon this calculation. Note that this result implies that variations in the relation between measured hours of work and the true labor input to the production due to "labor hoarding" are not equivalent in all respects to variations in effort, despite the fact that the two phenomena are sometimes treated as interchangeable.[36]

If we allow for variation in the degree to which the measured labor input provides inputs to current production (either due to labor hoarding or to effort variations), one could also, in principle, measure marginal cost by considering the cost of increasing output along that margin, holding fixed the measured labor input. Consideration of this issue would require modeling the cost of higher utilization of the labor input for production purposes. One case in which this does not involve factors other than those already considered here is if higher effort requires that labor be better compensated, owing to the existence of an effort-wage schedule $w(e)$ of the kind assumed by Sbordone (1996). In this case the marginal cost of increasing output by demanding increased effort results in an expression of the form (2.11), where now $\omega \equiv ew'(e)/w(e)$. If, at least in the steady state, the number of hours hired are such that the required level of effort is cost-minimizing, and that cost-minimizing effort level is unique, then (just as in our discussion above of a schedule specifying the wage as a function of hours per employee) the elasticity ω will be an increasing function of e, at least near the steady-state level of effort. The existence of procyclical effort variations would then, under this theory, mean that implied markup variations are more countercyclical than one would conclude if the effort variations were not taken into account.

This does not contradict the conclusion of the paragraph before last. For in a model like Sbordone's, effort variations should never be used by a firm, in the absence of adjustment costs for hours or employment (or some other reason for increasing marginal costs associated with increases in the measured labor input, such as monopsony power). In the presence, say, of adjustment costs, consideration of the marginal cost of increasing output through an increase in the labor input leads to (2.12), rather than to (2.5); this is consistent with the

above analysis, since a cost-minimizing choice of the level of effort to demand requires that

$$\omega(e) = \Omega \tag{2.19}$$

at all times. It is true (as argued two paragraphs ago) that variable effort requires no change in the derivation of (2.12). But observation of procyclical effort variations could provide indirect evidence of the existence of adjustment costs, and hence of procyclical variation in the factor Ω.

A further complication arises if the cost to the firm of demanding greater effort does not consist of higher current wages. Bils and Kahn (1996), for example, assume that there exists a schedule $w(e)$ indicating the effective cost to the firm of demanding different possible effort levels, but that the wage that is actually paid is independent of the current choice of e, due to the existence of an implicit contract between firm and worker of the form considered in Hall (1980). They thus suppose that the current wage equals $w(e^*)$, where e^* is the "normal" (or steady-state) level of effort. In this case, (2.12) should actually be

$$\mu = \Omega^{-1}\eta_H s_H^{-1}\frac{w(e^*)}{w(e)}. \tag{2.20}$$

If effort variations are procyclical, the factor $w(e)/w(e^*)$ is procyclical, and so this additional correction makes implied real marginal costs even more procyclical. In their empirical work they relate $w(e)/w(e^*)$ to variations in the energy consumption per unit of capital and show that this correction makes marginal cost significantly procyclical in four of the six industries they study. Interestingly, these four industries have countercyclical marginal costs when they ignore variations in the cost of labor that result from variations in effort.

Variable Utilization of Capital.

It is sometimes argued that the degree of utilization of firms' capital stock is procyclical as well, and that the production function is therefore properly a function of "effective" capital inputs that do not coincide with the measured value of firms' capital stocks. If by this one means that firms can produce more from given machines when more labor is used along with them, then it is not clear that "variable utilization" means anything that is not

already reflected in a production function of the form (2.1). Suppose, however, that it is possible for a firm to vary the degree of utilization of its capital stock other than by simply increasing its labor-to-capital ratio, and that the production function is actually of the form $Y = F(u_K K, zH)$, where u_K measures the degree of utilization of the capital stock K. Even so, the derivation of equation (2.5) is unaffected (and the same is true of subsequent variations on that equation, such as (2.8), (2.11), (2.12) and (2.18). The reason is that variation in capital utilization has no consequences for those calculations different from the consequences of time-variation in the capital stock itself. It is simply necessary to define y in (2.6) by y/u_K. In the case of an isoelastic production function (2.3), the methods of calculating implied markup variations we discussed above do not need to be modified at all.

Variable capital utilization matters in a more subtle way if one assumes that capital utilization depends upon aspects of the firm's labor input decisions other than the total labor input H. For example, Bils and Cho (1994) argue that capital utilization should be an increasing function of the number of hours worked *per employee*; the idea being that if workers remain on the shop floor for a longer number of hours each week, the capital stock is used for more hours as well (increasing the effective capital inputs used in production), whereas a mere increase in the number of employees, with no change in the length of their work-week, does not change the effective capital inputs used in production.[37] Under this hypothesis, the aggregate production function is given by $Y = F(u_K(h)K, zhN)$. This modification again has no effect upon the validity of the derivation of (2.12) from a consideration of the cost of increasing output by varying employment, holding hours per employee fixed (except, again, for the modification of (2.6)). Thus (2.15) becomes

$$\hat{\mu} = a\hat{y} - a\lambda\hat{h} - s\hat{_H} - \hat{\Omega}, \tag{2.21}$$

where λ is the elasticity of u_K with respect to h, while (2.16) is unchanged. If one assumes $a = 0$ (as Bils (1987) does), this would mean no change in the implied markup variations obtained using this method (which, as we have argued, is equivalent to Bils' "second method").[38]

Assuming that u_K depends upon h *does* affect our calculation of the cost of increasing

output by increasing hours per employee. In particular, (2.11) must instead be replaced by

$$\mu = \omega^{-1}(\eta_H + \lambda\eta_K)s_H^{-1}, \tag{2.22}$$

where η_K is the elasticity of output with respect to the effective capital input. However, while the presence of $\lambda > 0$ in (2.20) is of considerable importance for one's estimate of the *average* level of the markup (it increases it), it has less dramatic consequences for implied markup *fluctuations*. In the Cobb-Douglas case, η_H and η_K are both constants, and implied percentage variations in markups are independent of the assumed size of λ. Thus this issue has no effect upon the computations of Bils (1987).

If we maintain the assumption of constant returns but depart from the Cobb-Douglas case by supposing that η_H is countercyclical (because $\epsilon_{KH} < 1$), then allowance for $0 < \lambda \leq 1$ makes the factor $\eta_H + \lambda\eta_K$ less countercyclical. This occurs for two reasons; first, the factor $\eta_H + \lambda\eta_K$ decreases less with decreases in η_H (and in the limit of $\lambda = 1$, it becomes a constant), and second, the factor y/u_K (upon which η_H depends) is again less procyclical. Nonetheless, even if we assume that all countercyclical variation in this factor is eliminated, implied markup variations will still be as strongly countercyclical as they would be with a Cobb-Douglas production function.

To sum up, there are a number of reasons why the simple ratio of price to unit labor cost is likely to give an imprecise measure of cyclical variations in the markup. As it happens, many of the more obvious corrections to this measure tend to make implied markups more countercyclical than is that simple measure. Once at least some of these corrections are taken account of, one may easily conclude that markups vary countercyclically, as is found by Bils (1987) and Rotemberg and Woodford (1991).

2.3 Alternative Measures of Real Marginal Cost

Our discussion in sections 2.1 and 2.2 has considered for the most part a single approach to measuring real marginal cost (or equivalently, the markup), which considers the cost of increasing output through an increase in the labor input. However, as we have noted, if

firms are minimizing cost, the measures of real marginal cost that one would obtain from consideration of *each* of the margins along which it is possible to increase output should move together; thus each may provide, at least in principle, an independent measure of cyclical variations in markups. While cyclical variation in the labor input is clearly important, cyclical variations in other aspects of firms' production processes are observed as well. We turn now to the implications of some of these for the behavior of real marginal cost.

Intermediate inputs.

Intermediate input use (energy and materials) is also highly cyclical. Insofar as the production technology does not require these to be used in fixed proportions with primary inputs (and Basu, 1995, presents evidence that in U.S. manufacturing industries it does not), this margin may be used to compute an alternative measure of real marginal cost. Consideration of this margin is especially attractive insofar as these inputs are not plausibly subject to the kind of adjustment costs involved in varying the labor input (Basu and Kimball, 1994), so that at least some of the measurement problems taken up in section 2.2 can be avoided.

Suppose again that gross output Q is given by a production function $Q(V, M)$, where V is an aggregate of primary inputs, and M represents materials inputs. Then, considering the marginal cost of increasing output by increasing materials inputs alone yields the measure

$$\mu = \frac{PQ_M(V, M)}{P_M}. \tag{2.23}$$

by analogy with (2.2). (Note that in (2.23), μ refers to the "gross-output" markup which we called μ^G in (2.10). Also note that P now refers to the price of the firm's product, and not a value-added price index as before.) Under the assumption that Q exhibits constant returns to scale,[39] Q_M is a decreasing function of M/V, or equivalently of the materials ratio $m \equiv M/Q$. In this case, log-linearization of (2.23) yields

$$\hat{\mu} = f\hat{m} - \hat{p}_M, \tag{2.24}$$

where $f < 0$ is the elasticity of Q_M with respect to m, and $\hat{p}_M$ indicates percentage fluctuations in the relative price of materials $p_M \equiv P_M/P$.

Both terms on the right-hand side of (2.24) provide evidence that markups vary countercyclically. Basu (1995) shows that intermediate inputs (energy and materials) rise relative to the value of output in expansions, at least when these are not due to technology shocks.[40] Basu furthermore assumes that p_M is equal to one because he views materials inputs as indistinguishable from final output. Under this assumption, the increase of m in booms immediately implies that markups are countercyclical.

In fact, however, goods can be ranked to some extent by "stage of processing"; all goods are not used as both final goods and intermediate inputs of other sectors to the same extent. And it has long been observed that the prices of raw materials rise relative to those of finished goods in business expansions, and fall relative to those of finished goods in contractions (*e.g.*, Mills, 1936; Means *et al.*, 1939). Murphy, Shleifer, and Vishny (1989) show that this pattern holds up consistently both when they consider broad categories of goods grouped by stage of processing, and when they consider particular commodities that are important inputs in the production of other particular goods. Hence it would seem that for the typical industry, p_M is a procyclical variable. Because of (2.24), this would itself be evidence of countercyclical markup variation, even if one regarded Q_M as acyclical. The combination of these two facts clearly supports the view that real marginal costs are procyclical, and hence that markups are countercyclical.

Note that in the case that the production function $Q(V, M)$ is isoelastic in M, (2.23) implies that μ should be inversely proportional to the share of materials costs in the value of gross output, $s_M \equiv p_M m$. Thus in this case the materials share would directly provide a suitable proxy for variations in real marginal cost, just as in our previous discussion of the labor share. However, this specification (implying a unit elasticity of substitution between intermediate and primary inputs) is hardly plausible. Rotemberg and Woodford (1996b) estimate elasticities of substitution for 20 two-digit manufacturing sectors, and find an average elasticity less than 0.7. Basu's (1995) estimate of the response of m to changes in the relative price of primary and intermediate inputs suggests an elasticity of substitution half that size.[41] Thus it seems most likely that instead $f < -1$ in (2.24). If the materials

ratio m is procyclical as found by Basu, it follows that real marginal costs are actually more procyclical than is indicated by the materials share alone.

A related measure is used by Domowitz, Hubbard and Petersen (1986), who measure "price-cost margins" defined as the ratio of price to "average variable cost". They measure this as a the ratio of industry revenues to the sum of labor and materials costs, which is to say, as the reciprocal of the sum of the labor and materials shares. This should correspond to the markup as we have defined it only under relatively special circumstances. If the production function is isoelastic in *both* labor inputs and materials inputs, then real marginal cost is proportional to the labor share (as explained in section 2.1), and also proportional to the materials share (as explained in the previous paragraph). It then follows that these two shares should move in exact proportion to one another, and hence that their *sum* is a multiple of real marginal cost as well. Domowitz *et al.* report that this sum is somewhat countercyclical for most industries, and as a result they conclude that price-cost margins are generally procyclical. However, the conditions under which this measure should correspond to variations in the markup of price over marginal cost are quite restrictive, since they include all of the conditions required for the labor share to be a valid measure of real marginal cost, *and* all of those required for the materials share to be a valid measure. We have reviewed in section 2.2 a number of reasons why the labor share is probably less procyclical than is real marginal costs. Similar considerations apply in the case of the materials share, although the likely quantitative importance of the various corrections is different in the two cases; in the case of materials, the elasticity of substitution below unity is probably a more important correction, while adjustment costs are probably much less important. Nonetheless, one must conclude, as with our previous discussion of the labor share alone, that real marginal cost is likely to be significantly more procyclical than is indicated by the Domowitz *et al.* measure of "average variable cost".[42]

Inventory fluctuations.

Another margin along which firms may increase the quantity of goods available for sale

in a given period is by drawing down inventories of finished goods. For a cost-minimizing firm, the marginal cost of drawing down inventories must at all times equal the marginal cost of additional production, and thus measurement of the costs of reduced inventories provides another potential (indirect) measure of the behavior of marginal cost.

The following simple framework will clarify what is involved in such an analysis. Inventories at the end of period t, I_{t+1}, equal $I_t + Q_t - S_t$, where Q_t is production at t and S_t are sales at t. It is thus possible for a firm to keep its path of sales (and hence revenues) unchanged, increasing production and inventories at time t by one unit while reducing production by one unit at time $t+1$. If the firm's production and inventory-holding plan is optimal, such a marginal deviation should not affect the present value of its profits. For the typical firm, the proposed deviation raises nominal costs by the marginal cost of production at t, c_t, while lowering them by the present value of the marginal cost of production at $t+1$, and also raising profits by the marginal benefit of having an additional unit of inventory at the end of t. Denoting the real value of this latter marginal benefit by $b(I_t, Z_t)$, where Z_t denotes other state variables at date t that may affect this benefit, we have

$$P_t b(I_t, Z_t) + E_t\{R_{t,t+1} c_{t+1}\} = c_t$$

as a first-order condition for optimal inventory accumulation by the firm, where P_t is the general price level at date t (not necessarily the price of the firm's output), and $R_{t,t+1}$ is a stochastic discount factor for nominal income streams. This may equivalently be written

$$b(I_t, Z_t) = (c_t/P_t) - E_t \rho_{t,t+1}(c_{t+1}/P_{t+1}), \tag{2.25}$$

where now $\rho_{t,t+1}$ is the corresponding discount factor for real income streams. Given an assumption about the form of the marginal benefit function $b(I, Z)$, observed inventory accumulation then provides evidence about real marginal costs in an industry – more precisely, about the expected *rate of change* in real marginal costs.

The early studies in this literature (*e.g.*, Eichenbaum, 1989; Ramey, 1991) have tended to conclude that real marginal cost is countercyclical. The reason is that they assume that

the marginal benefit of additional inventories should be decreasing in the level of inventories (or equivalently, that the marginal cost of holding additional inventories is increasing); the finding that inventories are relatively high in booms then implies that b is low, from which the authors conclude that real marginal costs must be temporarily low.[43] Eichenbaum interprets the countercyclical variation in real marginal costs as indicating that output fluctuations are driven by cost shocks, while Ramey stresses the possibility that increasing returns to scale could be so pervasive that marginal cost could actually be lower in booms. Regardless of the explanation, if the finding of countercyclical real marginal costs is true for the typical sector, it would follow that markups in the typical sector must be procyclical. This is indeed the conclusion reached by Kollman (1996).

Bils and Kahn (1996) argue, instead, that real marginal cost is procyclical in each of the six production-for-stock industries that they investigate. The differing conclusion hinges upon a different conclusion about cyclical variation in the marginal benefits of additional inventories. They begin by observing that inventory-to-sales ratios do not vary secularly. This suggests that the function b is homogeneous of degree zero in inventories and sales; specifically, they propose that b is a decreasing function, not of I alone, but of I/S.[44] A similar conclusion follows from noticing that inventory-to-sales ratios are fairly constant across different models of automobiles at a given point in time, even though these models differ dramatically in the volume of their sales.

But this implies that b is actually *higher* in booms. The reason is that, as Bils and Kahn show, the ratio of inventories to sales is strongly *countercyclical*; while inventories rise in booms, they rise by less than do sales. Thus, the marginal value of inventories must be high in booms and, as a result, booms are periods where real marginal costs are temporarily high.

This conclusion is consistent both with the traditional view that diminishing returns result in increasing marginal costs, and with the view that business cycles are not primarily due to shifts in industry cost curves. As noted earlier, Bils and Kahn also show that their inventory-based measures of real marginal cost covary reasonably closely with a wage-based measure of the kind discussed above, once one corrects the labor cost measure for the exis-

tence of procyclical work effort as in (2.20). If their conclusion holds for the typical industry, and not just the six that they consider, it would have to imply countercyclical markup variations.[45]

Variation in the Capital Stock.

A final way in which output can be increased is by increasing the stock of capital.[46] Thus

$$\mu = \frac{PF_K(K, zH)}{E(r)}, \tag{2.26}$$

where $E(r)$ is the expected cost of increasing the capital stock at t by one unit while leaving future levels of the capital stock unchanged. Assuming that the capital stock at t can actually be changed at t but also letting there be adjustment costs, r_t equals

$$P_{K,t} + c_{I,t} - R_{t,t+1}(1-\delta)(P_{K,t+1} + c_{I,t+1})$$

where $P_{K,t}$ is the purchase price of capital at t, $c_{I,t}$ is the adjustment cost associated with increasing the capital stock at t by one unit, δ is the depreciation rate. It then becomes possible to measure changes in μ by differentiating (2.26). This is somewhat more complicated than the computation of marginal cost using either labor or materials because the rental rate of capital r cannot be observed directly; it must be inferred from a parametric specification for c_I.

A related exercise is carried out by Galeotti and Schiantarelli (1995). After specifying a functional form for c_I and making a homogeneity assumption regarding F, they estimate (2.26) by allowing μ to be a linear function of both the level of output and of expected changes in output. Their conclusion is that markups fall when the level of output is unusually high and when the expected change in output is unusually low. As we discuss further in section 3, this second implication is consistent with certain models of implicit collusion.

2.4 The Response of Factor Prices to Aggregate Shocks

Thus far we have discussed only the overall pattern of cyclical fluctuations in markups. Here we take up instead the degree to which markup variations play a role in the observed response

of the economy to particular categories of aggregate shocks. We are especially interested in shocks that can be identified in the data, that are known to be non-technological in character and that are thus presumptively statistically independent of variations in the rate of technical progress.[47] These cases are of particular interest because we can then exclude the hypothesis of shifts in supply costs due to changes in technology as an explanation for the observed response of output and employment This allows us to make judgments about the nature of markup variations in response to such shocks that are less dependent upon special assumptions about the form of the production function than has been true above (where such assumptions were necessary in order to control for variable growth in technology).

In particular, in the case of a variation in economic activity as a result of a non-technological disturbance, if markups do not vary, then real wages should move *countercyclically*. In our basic model, this is a direct implication of (2.2), under the assumption of a diminishing marginal product of labor.[48] For in the short run, the capital stock is a predetermined state variable, and so increases in output can occur if and only if hours worked increase, as a result of which the marginal product of labor must decrease; this then requires a corresponding decrease in the real wage, in order to satisfy (2.2). In the case of such a shock, then, the absence of countercyclical real wage movement is itself evidence of countercyclical markup variation.

Before turning to the evidence, it is worth noting that the inference that procyclical (or even acyclical) real wages in response to these shocks imply countercyclical markups is robust to a number of types of extension of the simple model that leads to (2.2). For example, the presence of overhead labor makes no (qualitative) difference for our conclusion, since the marginal product of labor should still be decreasing in the number of hours worked. A marginal wage not equal to the average wage also leads to essentially the same conclusion. If, in particular, we assume that the firm's wage bill is a nonlinear function of the form $W(H) = w_0 v(H)$, where the function $v(H)$ is time-invariant though the scale factor w_0 may be time-varying,[49] then $\omega(H)$, the ratio of the marginal to the average wage, is a time-invariant function. Since the denominator of (2.2) should actually be the *marginal* wage,

when the two differ, our reasoning above actually implies that $\mu\omega$ must be countercyclical. But as we have explained above, $\omega(H)$ is likely to be an increasing function (if it is not constant), so that μ should vary even more countercyclically than does the product $\mu\omega$ (which equals the ratio of the marginal product of labor to the average wage). If there are convex costs of adjusting the labor input, one similarly concludes that $\mu\Omega$ must be countercyclical. But since the factor Ω (defined in (2.13)) will generally vary procyclically, this is again simply a reason to infer an even stronger degree of countercyclical variation in markups than is suggested by (2.2).

If there is labor hoarding, it can still be inferred in the case of an increase in output due to a non-technological disturbance that $H - H_m$ must have increased; and then, if real wages do not fall, (3.8) implies that markups must have declined. In the case of variable capital utilization, the situation is more complicated. Condition (2.2) generalizes to

$$\mu = \frac{PzF_H(u_K K, zH)}{W}. \tag{2.27}$$

If we assume as above that F is homogeneous degree one, F_H is a decreasing function of $zH/u_K K$. But the mere fact that output and the labor input increase will not settle the question whether the ratio of labor inputs to *effective* capital inputs, $zH/u_K K$, has increased or not. Hence it may not be clear that the marginal product of labor must decline in booms.

Suppose, however, that the cost of higher capital utilization consists of a faster rate of depreciation of the capital stock. Let the rate of depreciation be given by $\delta(u_K)$, and let $\tilde{V}(K')$ denote the value to the firm of having an undepreciated capital stock of K' at the end of the period. The usual assumption of diminishing returns makes it natural to suppose that δ should be an increasing, convex function, while $\tilde{V}$ should be an increasing, concave function.[50] Then if we consider the marginal cost of increasing output solely by increasing the rate of utilization of the capital stock, we obtain the additional relation

$$\mu = \frac{F_K(u_K K, zH)}{\tilde{V}'((1 - \delta(u_K))K)\delta'(u_K)}. \tag{2.28}$$

Now if $zH/u_K K$ decreases when output expands, it follows that F_K declines. Furthermore, this requires an increase in u_K, so that, under our convexity assumptions, both $\tilde{V}'$ and δ'

must increase. Thus (2.28) unambiguously requires the markup to decrease. Alternatively, if $zH/u_K K$ increases, F_H declines, and then, if there is no decline the real wage, (2.27) requires a decline in the markup. Thus under either hypothesis, markup variations must be countercyclical, if real wages are not.[51]

We turn now to the question of whether expansions in economic activity associated with non-technological disturbances are accompanied by declines in real wages. There are three important examples of identified non-technological disturbances that are often used in the literature. These are variations in military purchases, variations in the world oil price, and monetary policy shocks identified using "structural VAR" methods. At least in the U.S., the level of real military purchases has exhibited noticeable variation over the post-WWII period (as a result of the Korean conflict, Vietnam, and the Reagan-era military build-up). The causes of these variations are known to have had to do with political events that have no obvious connection with technical progress. (We consider military purchases rather than a broader category of government purchases exactly because this claim of exogeneity is more easily defended in the case of military purchases.) Similarly, the world oil price has been far from stable over that period (the two major "oil shocks" of the 1970's being only the most dramatic examples of variation in the rate of increase in oil prices), and again the reasons for these variations, at least through the 1970's, are known to have been largely external to the U.S. economy (and to have had much to do with political dynamics within the OPEC cartel).[52] In the case of monetary policy shocks, the identification of a time series for exogenous disturbances is much less straightforward (since the Federal funds rate obviously responds to changes in economic conditions, including real activity and employment, as a result of the kind of policies that the Federal Reserve implements). However, an extensive literature has addressed the issue of the econometric identification of exogenous changes in monetary policy,[53] and we may therefore consider the estimated responses to these identified disturbances. In each of the three cases, the variable in question is found to be associated with variations in real activity, and these effects are (at least qualitatively) consistent with economic theory, so that it is not incredible to suppose that

the observed correlation represents a genuine causal relation.

We turn now to econometric studies of the responses to such shocks, using relatively unrestricted VAR models of the aggregate time series in question. Rotemberg and Woodford (1989) show that increases in real military purchases raise private value added, hours worked in private establishments and wages deflated by the relevant value added deflator. Ramey and Shapiro (1997) show that the effect on this real wage is different when revised NIPA data are used and that, with revised data, this real wage actually falls slightly. They argue that this response can be reconciled with a two-sector constant markup model. Whether a one-sector competitive model can be reconciled with their evidence remains an open question.

Christiano, Eichenbaum and Evans (1996) show, using a structural VAR model to identify monetary policy shocks, that output and real wages both decline in response to the increases in interest rates that are associated with monetary tightening. This again suggests that the contraction in output is associated with an increase in markups. An increase in the federal funds rate by one percent that leads to a 0.4 percent reduction in output reduces real wages by about 0.1 percent. If one supposes that hours fall by about the same percent as output, the effective increase in the markup is about 0.2 percent.

Rotemberg and Woodford (1996) look instead at the response of the U.S. economy to oil price increases. They show that during the pre-1980 OPEC period, such increases lowered private value added together with real wages. Specifically, a one percent unexpected increase in oil prices is shown to lead to a reduction of private value added by about a quarter of a percent after a year and a half, and to a reduction of the real wage (hourly earnings in manufacturing deflated by the private value-added deflator) by about 0.1 percent, with a similar time lag. This combination of responses again suggests that markups increase, especially during the second year following the shock.

The inference is, however, less straightforward in this case; for one might think that an increase in oil prices should have an effect similar to that of a negative technology shock, even if it does not represent an actual change in technology. In fact, Rotemberg and Woodford show that this is not so. Let us assume again the sort of separable utility function used

to derive (2.23), but now interpret the intermediate input "M" as energy. In this case, consideration of the marginal cost of increasing output by increasing labor inputs yields

$$\mu = \frac{PQ_V(V,M)V_H(K,zH)}{w}. \tag{2.29}$$

Comparison of (2.29) with (2.23) allows us to write a relation similar in form to (2.2),

$$\mu = \frac{\tilde{P}V_H(K,zH)}{W}, \tag{2.30}$$

where the price index $\tilde{P}$ is defined by

$$\tilde{P} \equiv \frac{PY - \mu P_M M}{V(K,zH)}. \tag{2.31}$$

Thus if we deflate the wage by the proper price index $\tilde{P}$, it is equally true of an energy price change that a decrease in labor demand must be associated with an increase in the real wage, unless the markup rises. (Note that the situation is quite different in the case of a true technology shock, since the relation (2.30) is shifted by a change in z.)

Under the assumption of perfect competition ($\mu = 1$), the price index defined in (2.31) is just the ideal (Divisia) value-added deflator. Thus a competitive model would require the value-added-deflated real wage to rise following an oil shock, if employment declines;[54] and the observation that real wages (in this sense) decline would suffice to contradict the hypothesis of perfect competition. The results of Rotemberg and Woodford do not quite establish this; first, because their private-value-added deflator is not precisely the ideal deflator, but more importantly, because their measure of private value added includes the U.S. energy sector, whereas the above calculations refer to the output of non-energy producers (that use energy as an input). Still, because the energy sector is small, even the latter correction is not too important quantitatively; and Rotemberg and Woodford show, by numerical solution of a calibrated model under the assumption of perfect competition, that while small simultaneous declines in their measure of output and of the real wage would be possible under competition, the implied declines are much smaller than the observed ones.[55]

Similar reasoning allows us to consider as well the consequences of changes in the relative price of intermediate inputs other than energy. We ignored materials inputs in our discussion

above of the inferences that may be drawn from the response of real wages to identified shocks. As before, however, equation (2.2) (and similarly (2.27)) can be interpreted as referring equally to a production technology in which materials inputs are used in fixed proportions with an aggregate of primary inputs, under the further assumption that the relative price of materials is always one, because materials and final goods are the same goods. But the relative prices of goods differing by "stage of processing" do vary, and so a more adequate analysis must take account of this. When one does so, however, one obtains (2.30) instead of (2.2). It is still the case that the failure of real wages to rise in the case of a non-technological disturbance that contracts labor demand indicates that markups must rise, as long as the real wage in question is $w/\tilde{P}$.

What, instead, if one observes only the behavior of w/P? Then the failure of this real wage to rise might, in principle, be explained by a decline in $\tilde{P}/P$, consistent with a hypothesis of constant (or even procyclical) markups. However (referring again to (2.29)), this would require a decline in $Q_V(V, M)$. Under the assumption that Q is homogeneous degree one, this in turn would require a decline in M/V, hence an increase in $Q_M(V, M)$. If markups are constant or actually decreasing, this would then require an increase in the relative price of materials, P_M/P, by (2.23). Thus we can extend our previous argument to state that if one observes that *neither* w/P nor P_M/P increases in the case of a non-technological disturbance that leads to reduced labor demand, one can infer that markups must increase. In fact, Clark (1996) shows, in the case of a structural VAR identification of monetary policy disturbances similar to that of Christiano *et al.*, that a monetary tightening is followed by increases in the price of final goods relative to intermediate goods and raw materials. This, combined with the evidence of Christiano *et al.* regarding real wage responses, suggests that a monetary tightening involves an increase in markups.

A possible alternative explanation of declines in real wages and the relative price of materials inputs at the same time as a contraction of output and employment is an increase in some other component of firms' marginal supply cost. Christiano *et al.* propose that an increase in financing costs may be the explanation of their findings.[56] As they show, in a

model where firms require bank credit to finance their wage bill, the interest rate that must be paid on such loans also contributes to the marginal cost of production; and it is possible to explain the effects of a monetary tightening, without the hypothesis of markup variation, as being due to an increase in marginal cost due to an increase in the cost of credit. But while this is a theoretical possibility, it is unclear how large a contribution financing costs make to marginal costs of production in reality.[57] This matter deserves empirical study in order to allow a proper quantitative evaluation of this hypothesis.

2.5 Cross-Sectional Differences in Markup Variation

In this subsection we survey the relatively scant literature that investigates whether markups are more countercyclical in industries where it is more plausible *a priori* that competition is imperfect. This issue is of some importance because countercyclical markups are less plausible in industries where there is little market power. For markups below one imply that the firm can increase its current profits by rationing consumers to the point at which marginal cost is no higher than the firm's price. But if markups never fall below one, there is little room for markup variation unless average markups are somewhat above one. In addition, the theoretical explanations we present for countercyclical markups in section 3 all involve imperfect competition. A consideration of whether the measures of markup variation that we have proposed imply that markup variation is associated with industries with market power is thus a check on the plausibility of our interpretation of these statistics. Quite apart from this, evidence on comparative markup variability across industries can shed light upon the adequacy of alternative models of the sources of markup variation.

The most straightforward way of addressing this issue is to compute markups for each sector using the methods discussed in section 2, and compare the resulting markup movements to output movements. In Rotemberg and Woodford (1991), we carry out this exercise for two-digit U.S. data, treating each of these sectors as having a different level of average markups and using Hall's (1988) method for measuring the average markup in each sector.[58] We show that the resulting markups are more negatively correlated with GNP in sectors

whose eight-digit SIC sector has a higher average four-firm concentration ratio. Thus, assuming this concentration is a good measure of market power, these results suggest that sectors with more imperfect competition tend to have more countercyclical markups.

One source of this result is that, as shown earlier by Rotemberg and Saloner (1986), real product wages W_i/P_i are more positively correlated with GNP, and even with industry employment, in more concentrated industries. By itself, this is not sufficient to demonstrate that markups are more countercyclical since zF_H could be more procyclical in these sectors. However, the analysis of Rotemberg and Woodford (1991) suggests that this is not the explanation for the more procyclical real product wages in more concentrated sectors.

As we discussed earlier, Domowitz, Hubbard and Peterson (1986) measure markup changes by the ratio of the industry price relative to a measure of "average variable cost". They show that this ratio is more procyclical in industries where the average ratio of revenues to materials and labor costs is larger, and see this as suggesting that markups are actually more procyclical in less competitive industries. As we already mentioned, this method for measuring markup variation imparts a procyclical bias for a variety of reasons. This bias should be greater in industries with larger fixed (or overhead) costs (because of equation (2.8)), and these are likely to be the more concentrated industries. In addition, the ratio of revenues to labor and materials costs is a poor proxy for the extent to which a sector departs from perfect competition, because this indicator is high in industries that are capital-intensive, regardless of the existence of market power in their product markets.

Domowitz, Hubbard and Petersen (1987) use a different method for measuring industry markup variations and obtain rather different results. In particular, they run regressions of changes in an industry's price on changes in labor and materials cost as well as a measure of capacity utilization. Using this technique, they show that prices are more countercyclical, *i.e.*, fall more when capacity utilization becomes low, in industries with higher average ratios of revenues to materials and labor costs. If the relation between capacity utilization and marginal cost were the same across industries, and if one accepted their method for deciding which industries are less competitive, their study would thus show that markups are more

countercyclical in less competitive industries.

3 Implications of Markup Variations for Business Fluctuations

In this section, we study whether it is empirically plausible to assign a large role on markup fluctuations in explaining business fluctuations. We first take up two related aspects of the observed cyclical variation in the relation between input costs and the value of output, that are sometimes taken to provide *prima facie* evidence for the importance of cost shifts (as opposed to markup changes) as the source of fluctuations in activity. These are the well-known procyclical variation in productivity and in profits. We show that these procyclical variations contain very little information on the importance of markup changes because markup variations induce such procyclical responses.

We next take up a more ambitious attempt at gauging the role of markup fluctuations in inducing cyclical fluctuations in economic activity. In particular, we study the extent to which the markup changes that we measured in sections 2.1 and 2.2 lead to output fluctuations. Any change in output that differs from that which is being induced by changes in markups ought naturally to be viewed as being due to a shift in real marginal costs (for a given level of output). Thus, this approach allows us to decompose output changes into those due to markup changes and those due to shifts in the marginal cost curve. What makes this decomposition particularly revealing is that, under the hypothesis that markups are constant *all output fluctuations are due to shifts in real marginal costs.*

3.1 Explaining Cyclical Variation in Productivity and Profits

Cyclical Productivity.

Standard measures of growth in total factor productivity (the "Solow residual" and variants) are highly positively correlated with growth in output and this fact is cited in the real business cycle literature (*e.g.*, Plosser (1989)) as an important piece of evidence in favor of the hypothesis that business cycles are largely due to exogenous variations in the rate of

technical progress. It might seem that variations in economic activity due to changes in firms' markups (in the absence of any shift in the determinants of the real marginal cost schedule) should not be associated with such variations in productivity growth, and that the close association of output variations with variations in productivity growth therefore leaves little role for markup variations in the explanation of aggregate fluctuations – or at least, little role for disturbances that affect economic activity *primarily* through their effect upon markups rather than through their effect on production costs.

In fact, however, there are a number of reasons why variations in markups should be expected to produce fluctuations in measured total factor productivity growth, that are strongly and positively correlated with the associated fluctuations in output growth. Thus observation of procyclical productivity growth does not in itself provide any evidence that markup variations do not play a central role in accounting for observed aggregate fluctuations. (Of course, procyclical productivity is not in itself conclusive evidence of markup variation either, since other explanations remain possible. For this reason productivity variations are a less crucial statistic than those discussed in sections 2.1 and 2.2.)

One reason is simply that standard measures of total factor productivity growth may use incorrect measures of the elasticities of the production function with respect to factor inputs. If these elasticities are assigned values that are too small (in particular, the elasticity η_H with respect to the labor input), then spurious procyclical variation in total factor productivity growth will be found. As Hall (1988) notes, the Solow residual involves a biased estimate of just this kind, if firms have market power. Consider a production function of the form (2.1), where F is not necessarily homogeneous of degree 1. Differentiation yields

$$\hat{\gamma}_Y = \eta_K \hat{\gamma}_K + \eta_H(\hat{\gamma}_z + \hat{\gamma}_H). \tag{3.1}$$

As noted before, (2.2) implies that $\eta_H = \mu s_H$; similar reasoning (but considering the marginal cost of increasing output by increasing the quantity of capital used) similarly implies that $\eta_K = \mu s_K$. Thus under perfect competition (so that $\mu = 1$), the elasticities correspond simply to the factor shares, and a natural measure of technical progress is given by the

Solow residual

$$\epsilon^{Solow} \equiv \hat{\gamma}_Y - s_K\hat{\gamma}_K - s_H\hat{\gamma}_H.$$

More generally, however, substitution of (3.1) (with the elasticities replaced by μ times the corresponding factor income share) yields

$$\epsilon^{Solow} = \frac{\mu - 1}{\mu}\hat{\gamma}_Y + s_H\hat{\gamma}_z. \tag{3.2}$$

In the case of perfect competition, only the second term is present in (3.2), and the Solow residual measures growth in the technology factor z. But in the presence of market power ($\mu > 1$), increases in output will result in positive Solow residuals (and decreases in output, negative Solow residuals), even in the absence of any change in technology. In particular, output fluctuations due to changes in the markup will result in fluctuations in the Solow residual, closely correlated with output growth.

Hall (1990) points out that in the case that the production function exhibits constant returns to scale, this problem with the Solow residual can be eliminated by replacing the weights s_K, s_H by the shares of these factor costs in *total costs*, rather than their share in revenues. Thus he proposes a "cost-based productivity residual"

$$\epsilon^{Hall} \equiv \hat{\gamma}_Y - \tilde{s}_K\hat{\gamma}_K - \tilde{s}_H\hat{\gamma}_H,$$

where $\tilde{s}_H \equiv s_H/(s_K + s_H)$, and $\tilde{s}_K = 1 - \tilde{s}_H$. In terms of these factor shares, the production function elasticities are given by $\eta_H = \rho\tilde{s}_H$, $\eta_K = \rho\tilde{s}_K$, where $\rho = \eta_K + \eta_H$ is the index of returns to scale defined earlier. Similar manipulations as are used to derive (3.2) then yield

$$\epsilon^{Hall} = \frac{\rho - 1}{\rho}\hat{\gamma}_Y + \tilde{s}_H\hat{\gamma}_z. \tag{3.3}$$

Even if $\mu > 1$, as long as $\rho = 1$, Hall's "cost-based" residual will measure the growth in z. One can show, in fact, that this measure of productivity growth is procyclical to essentially the same degree as is the Solow residual.[59] But again this need not indicate true technical change. For if there are increasing returns to scale ($\rho > 1$), due for instance to the existence of overhead labor as discussed above, then increases in output will result in positive Solow

residuals even without any change in technology. This explanation for the existence of procyclical productivity in the absence of cyclical changes in technology is closely related to the previous one, since we have already indicated that (given the absence of significant pure profits) it is plausible to assume that μ and ρ are similar in magnitude.

The quantitative significance of either of these mechanisms depends upon how large a value one believes it is plausible to assign to μ or ρ. Hall (1988, 1990) argues that many U.S. industries are characterized by quite large values of these parameters. He obtains estimates of μ that exceed 1.5 for 20 of the 26 industries for which he estimates this parameter. Within his 23 manufacturing industries, 17 have estimates of μ above 1.5 while 16 have estimates of ρ that are in excess of 1.5. His evidence is simply that both productivity residuals are positively correlated with output movements, even those output movements that are associated with non-technological disturbances. In effect, he estimates the coefficients on the first terms on the right-hand sides of (3.2) and (3.3) by instrumental-variables regression in using military purchases, a dummy for the party of the U.S. President, and the price of oil as instruments for non-technological disturbances that affect output growth. However, even assuming that the correlations with these instruments are not accidental, this merely establishes that some part of the procyclical productivity variations that are observed are not due to fluctuations in true technical progress; since explanations exist that do not depend upon large degrees of market power or increasing returns, one cannot regard this as proving that μ and ρ are large.

A second possible mechanism is substitution of intermediate for primary inputs, as discussed by Basu (1995). Suppose that materials inputs are not used in fixed proportions, but instead that each firm's gross output Q is given by a production function $Q = Q(V, M)$, where M represents materials inputs and V is an index of primary input use (which we may call "economic value added"), and the function Q is differentiable, increasing, concave, and homogeneous degree one. As before, economic value added is given by a value-added production function $V = F(K, zH)$. Now consider a symmetric equilibrium in which the price of each firm's product is the same, and this common price is also the price of each firm's

materials inputs (which are the products of other firms). Consideration of the marginal cost of increasing output by increasing materials inputs alone then yields

$$\mu = Q_M(V, M). \tag{3.4}$$

Because of our homogeneity assumption, (3.4) can be solved for

$$M/V = m(\mu),$$

where m is a decreasing function. Then defining accounting value added as $Y \equiv Q - M$, one obtains

$$Y/V = Q(1, m(\mu)) - m(\mu). \tag{3.5}$$

Furthermore, as long as firms have some degree of market power ($\mu > 1$), (3.4) implies that $Q_M > 1$. Hence $Q(1, m) - m$ will be increasing in m, and (3.5) implies that Y/V, the ratio of measured value added to our index of "economic value added", will be a decreasing function of μ.

This implies that a decline in markups would result in an increase in measured value added Y even without any change in primary input use (and hence any change in V). This occurs due to the reduction of an inefficiency in which the existence of market power in firms' input markets leads to an insufficiently indirect pattern of production (too great a reliance upon primary as opposed to intermediate inputs). If one's measure of total factor productivity growth is based upon the growth in Y instead of V, then markup variations will result in variations in measured productivity growth that are unrelated to any change in technology. Since a markup decline should also increase the demand for primary factors of production such as labor, it will be associated with increases in employment, output, and total factor productivity – where the latter quantity increases because of the increase in Y/V even if the measurement problems stressed by Hall (relating to the accuracy of one's measure of the increase in V that can be attributed to the increase in primary factor use) are set aside.

The quantitative importance of such an effect depends upon two factors, the elasticity of the function m and the elasticity of the function $Q(1, m) - m$. The first depends upon

the degree to which intermediate inputs are substitutable for primary inputs. Basu (1995) establishes that materials inputs do not vary in exact proportion with an industry's gross output; in fact, he shows that output growth is associated with an increase in the relative use of intermediate inputs, just as (3.4) would predict in the case of an output increase due to a reduction in markups. The second elasticity depends upon the degree of market power in the steady state (*i.e.*, the value of μ around which we consider perturbations), because as noted above, the derivative of $Q(1,m) - m$ equals $\mu - 1$. Thus while Basu's mechanism is quite independent of Hall's, it too can only be significant insofar as the typical industry possesses a non-trivial degree of market power.

An alternative mechanism is "labor hoarding"; indeed, this is probably the most conventional explanation for procyclical productivity variations. If only $H - H_m$ hours are used for current production, but productivity growth is computed using total payroll hours H as a measure of the labor input, then a tendency of H_m to decline when $H - H_m$ increases will result in spurious procyclical variations in measured productivity growth. Furthermore, this is exactly what one should expect to happen, in the case of fluctuations in activity due to markup variations.

Suppose that the value to a firm (in units of current real profits that it is willing to forego) of employing H_m hours on maintenance (or other non-production) tasks is given by a function $v(H_m)$. It is natural to assume that this function is increasing but strictly concave. Then if the firm is a wage-taker, and there are no adjustment costs for varying total payroll hours H, the firm should choose to use labor for non-production tasks to the point at which

$$v'(H_m) = w/P. \tag{3.6}$$

Let us suppose furthermore that the real wage faced by each firm depends upon aggregate labor demand, according to a *wage-setting locus* of the form

$$w/P = \nu(H), \tag{3.7}$$

where ν is an increasing function.[60] Since v' is a decreasing function while ν is increasing, (3.6) and (3.7) imply that H and H_m should move inversely with one another, assuming that

the source of their changes is not a shift in either of the two schedules. Finally, allowing for labor allocated to non-production tasks requires us to rewrite (2.2) as

$$\mu = \frac{PzF_H(K, z(H - H_m))}{w}. \tag{3.8}$$

Substituting for H_m in the numerator the decreasing function of H just derived, and substituting for w in the denominator using (3.7), the right-hand side of (3.8) may be written as a decreasing function of H. It follows that a reduction in the markup (not associated with any change in the state of technology, the value of non-production work, or the wage-setting locus) will increase equilibrium H and reduce equilibrium H_m. The result will be an increase in output accompanied by an increase in measured total factor productivity.

If the firm faces a wage that increases with the total number of hours that it hires (due to monopsony power in the labor market, the overtime premium, or the like), then the resulting procyclical movements in measured productivity will be even greater. In this case, (3.6) becomes instead

$$v'(H_m) = \omega(H)w/P, \tag{3.9}$$

where $\omega(H)$ is the ratio of the marginal to the average wage, as in (2.11). We have earlier given several reasons why $\omega(H)$ would likely be an increasing function, at least near the steady-state level of hours. Hence the specification (3.9) makes the right-hand side an even more sharply increasing function of H than in the case of (3.6). Similarly, if there are convex costs of adjusting the total number of hours hired by the firm, (3.6) becomes instead

$$v'(H_m) = \Omega w/P, \tag{3.10}$$

where Ω is again the factor defined in (2.13). Again, this alternative specification makes the right-hand side an even more procyclical quantity than in the case of (3.6). Thus either modification of the basic model with labor hoarding implies even more strongly countercyclical movements in H_m, and as a result even more procyclical variation in measured productivity.

A related explanation for cyclical variation that results from markup variations in measured productivity is unmeasured variation in labor effort. If, as in the model of Sbordone

(1996), the cost of increased effort is an increase in the wage $w(e)$ that must be paid, and there are convex costs of varying hours, then the cost-minimizing level of effort for the firm is given by (2.19). As discussed earlier, this implies that effort should co-vary positively with fluctuations in hours (albeit with a lead), since the factor Ω will be procyclical with a lead, while the function $\omega(e)$ will be increasing in e. Furthermore, consideration of the marginal cost of increasing output by demanding increased effort implies that[61]

$$\mu = \frac{PzF_H(K, zeH)}{w'(e)}. \tag{3.11}$$

Since $w'(e)$ must be increasing in e (at least near the steady-state effort level, as a consequence of the second-order condition for minimization of the cost w/e of effective labor inputs), (3.11) requires that a reduction in markups result in an increase in eH (to lower the numerator), an increase in e (to increase the denominator), or both. Since e and H should co-vary positively as a consequence of (2.19), it follows that a temporary reduction of markups should be associated with temporary increases in effort, hours, and output. Countercyclical markup fluctuations would therefore give rise to procyclical variations in measured productivity.

Another related explanation is unmeasured variation in the degree of utilization of the capital stock. The argument in this case follows the same general lines. If markups fall, firms must choose production plans that result in their operating at a point of higher real marginal costs (which quite generally means more output). Cost-minimization implies that real marginal costs increase apace along *each* of the margins available to the firm. Thus if it is possible to independently vary capital utilization, the real marginal cost of increasing output along this margin must increase; under standard assumptions, this will mean more intensive utilization of the firm's capital. But the resulting procyclical variation in capital utilization will result in procyclical variation in measured productivity, even if there is no change in the rate of technical progress.

Similar conclusions are obtained when capital utilization is a function of hours worked per employee. Consider again the case in which there is an interior solution for hours because the wage schedule $W(h)$ is nonlinear in hours per employee, and in which hours per

employee nonetheless vary because of convex adjustment costs for employment. Then the cost-minimizing decision regarding hours per employee satisfies the first-order condition[62]

$$\Omega = \omega(h)\Big(\frac{\eta_H}{\eta_H + \lambda\eta_K}\Big). \tag{3.12}$$

If we assume both a Cobb-Douglas production function $Y = (u_K K)^{1-\alpha}(zhN)^{\alpha}$ and an isoelastic capital utilization function $u_K = h^{\lambda}$ with $0 < \lambda \leq 1$, the expression in parentheses is a constant, and (3.12) implies that hours per employee h must covary positively with Ω. This means that fluctuations in hours will accompany temporary fluctuations in employment (but with a lead). Furthermore, (2.22) implies that

$$\mu = \frac{[\alpha + \lambda(1-\alpha)]Pz^{\alpha}K^{1-\alpha}}{W'(h)h^{(1-\alpha)(1-\lambda)}N^{1-\alpha}}.$$

Thus a decline in μ must be accompanied by an increase in $W'(h)h^{(1-\alpha)(1-\lambda)}$ (hence an increase in h), an increase in $N^{1-\alpha}$ (hence an increase in N), or both. Since employment and hours must co-vary positively, there will be an increase in both. As a result, capital utilization will increase along with output and employment, again resulting in procyclical variation in measured productivity.

Cyclical Profits.

Business profits are also well-known to vary procyclically (*e.g.*, Hultgren, 1965); corporate profits after taxes have long been a component of the NBER's index of coincident business cycle indicators. This is sometimes thought to make it implausible that business expansions are associated with declines in markups, since reduced markups should lower profits. Indeed, Christiano, Eichenbaum and Evans (1996) report calculations intended to show that a model in which expansions are due to markup declines will almost inevitably make the counterfactual prediction that profits must decline when output expands.[63]

This implication is, however, less direct than it might at first seem. There are a number of reasons why profits might well rise when markups fall. Many of these have been introduced above as reasons why the inverse of the labor share need not move countercyclically to the same extent as the markup. The connection between these two issues is simple. The

cyclical variation in (real) profits is essentially determined by the cyclical variation in the amount by which the value of output exceeds the wage bill, $Y - (w/P)H$. (This is because the remaining deductions involved in the calculation of accounting profits, such as interest payments and depreciation allowances, are relatively less cyclical.) Now if the labor share wH/PY is *not* procyclical, it follows that when output increases, wH/P increases no more than proportionally to output, which surely means *less* in absolute magnitude, since labor compensation is on average only three-quarters of the value of output. Hence $Y - (w/P)H$ will increase. Thus any model that does not predict a procyclical labor share will *a fortiori* not predict countercyclical profits. And indeed, parameter values that imply procyclical variation in profits in response to markup variations are not hard to find.

Consider first our simplest model, in which firms pay the same wage regardless of the number of hours they hire, there are no adjustment costs, and the measured capital and labor inputs are all that matter for a firm's output. Then equilibrium output Y, hours H, and real wage w/P are determined by equations (2.1), (2.2), and (3.7), given the capital stock K, the state of technology z, and the markup μ. Let us consider the effects of markup variations, holding fixed the other two parameters (and the functions F and ν). If we neglect changes in interest and depreciation, the change in profits is given by

$$\begin{aligned} d\Pi &= d(Y - \nu H) = (zF_H - \nu)dH - Hd\nu \\ &= (\mu - 1 - \epsilon_\nu)\nu dH, \end{aligned} \tag{3.13}$$

where $\nu \equiv w/P$ is the real wage, and $\epsilon_\nu \equiv H\nu'/\nu$ is the elasticity of the wage-setting locus in (3.7). It follows that profits increase along with employment and output if and only if

$$\mu > 1 + \epsilon_\nu. \tag{3.14}$$

Now this is certainly possible; under the hypothesis of market power in the product market (which we require in order to suppose that markup variations are *possible*), $\mu > 1$, so it is simply necessary that ϵ_ν be small enough.

This may not, however, seem empirically plausible; essentially, Christiano *et al.* argue that it would require a greater degree of market power than is plausible for most U.S.

industries. Their proposed value for ϵ_ν, however (their "baseline" calculation assumes $\epsilon_\nu = 1$), is based not upon the observed degree of cyclicality of wages, but upon what they regard as a plausible specification of household preferences, given an interpretation of (3.7) as the labor supply schedule of representative household. In fact, the average wage is observed to be relatively acyclical, and even if this is a puzzle for the theory of labor supply, there is no reason to assume a stronger real wage response to increases in labor demand in calculating the effect on profits of an increase in output associated with a decline in markups. For example, Solon, Barsky and Parker (1994) find an elasticity of the average real wage with respect to hours worked of about 0.3;[64] thus an average markup in excess of 1.3 would suffice to account for procyclical profit variations. And again, this is the "value-added markup" that must exceed 1.3; for this, the typical supplier's markup need not be much more than ten percent.

In any event, procyclical profits do not require even as large an average markup as this, if we make the model more realistic, in any of the several ways discussed above. Consider first the possibility that the marginal wage paid by a firm varies with the number of hours that it hires, and not only with aggregate labor demand (as assumed above), due, for example, to monopsony power in the labor market. Let us write the firm's wage bill as $W(H^i; H)$, where H^i represents hours hired by firm i, and H represents aggregate hours hired. Then in a symmetric equilibrium, the average wage ν is given by $W(H;H)/H$, and the ratio ω of the marginal wage to the average wage is given by $HW_1(H;H)/W(H;H)$. In this case, equation (3.13) generalizes to

$$\begin{aligned} d\Pi &= d(Y - W(H;H)) &= (zF_H - W_1 - W_2)dH \\ &= (\mu - 1 - W_2/W_1)W_1 dH &= (\omega\mu - 1 - \epsilon_\nu)\nu dH, \end{aligned}$$

so that (3.14) becomes

$$\omega\mu > 1 + \epsilon_\nu. \tag{3.15}$$

Since, as explained earlier, there are a number of reasons for ω to be larger than one, the markup need not be as large as is required by (3.14) in order for profits to be procyclical.

If, for example, we assume that $\omega = 1.2$, as Bils (1987) estimates,[65] and $\epsilon_\nu = 0.3$, it suffices that $\mu = 1.1$ (which means a gross-output markup of 4 percent).

Alternatively, suppose that some labor is used for non-production purposes, as in our above discussion of "labor hoarding". Then (3.13) becomes instead

$$d\Pi = d(Y - \nu H) = zF_H(dH - dH_m) - \nu dH - Hd\nu$$

$$= (\theta\mu - 1 - \epsilon_\nu)\nu dH,$$

where θ denotes the derivative of labor used in production $H - H_m$ with respect to total labor H. Thus (3.14) becomes

$$\theta\mu > 1 + \epsilon_\nu. \tag{3.16}$$

If labor hoarding is countercyclical, $\theta > 1$, and (3.16) also requires a smaller markup than does (3.14). The findings of Fay and Medoff (1985), discussed above, would suggest a value of θ on the order of 1.4. This would be enough to satisfy (3.16) regardless of the size of the markup.

Similar results are obtained in the case of variable labor effort or variable capital utilization. The implied modification of (3.14) is largest if the costs of higher effort or capital utilization do not show up in accounting measures of current profits. For example, suppose that effective capital inputs are given by $u_K K$, where the utilization rate u_K is an independent margin upon which the firm can vary its production process, and suppose that the cost of higher utilization is faster depreciation of the capital stock (but that this is not reflected in the depreciation allowance used to compute accounting profits). As explained above, we should expect a decline in markups to be associated with a simultaneous increase in real marginal costs along each margin, so that firms choose to increase u_K at the same time that they choose to increase labor inputs per unit of capital. Let λ denote the elasticity of u_K with respect to H as a result of this cost-minimization on the part of firms.[66] Then (3.13) becomes instead

$$d\Pi = d(Y - \nu H) = zF_H dH + KF_K du_K - \nu dH - Hd\nu$$

$$= (\mu + \eta_K s_H^{-1}\lambda - 1 - \epsilon_\nu)\nu dH,$$

$$= \Big[\Big(\frac{\eta_H + \lambda\eta_K}{\eta_H}\Big)\mu - 1 - \epsilon_\nu\Big]\nu dH,$$

and (3.14) again takes the form (3.16), where now $\theta \equiv (\eta_H + \lambda\eta_K)/\eta_H$. If capital utilization and hours co-vary positively (as we have argued, and as is needed in order to interpret procyclical productivity variations as due to cyclical variation in capital utilization), then $\theta > 1$, and again a smaller markup than is indicated by (3.14) will suffice for procyclical profits. If, for example, $\lambda = 1$, as argued by Bils and Cho (1994), then $\theta > 1.3$, and (3.16) is satisfied no matter how small the average markup may be.

3.2 Identifying the Output Fluctuations Due to Markup Variation

We now describe the consequences of alternative measures of marginal costs for one's view of the sources of aggregate fluctuations. We propose to decompose the log of real GDP y_t as

$$y_t = y_t^* + \hat{y}_t^\mu, \tag{3.17}$$

where the first term represents the level of output that is warranted by shifts in the real marginal cost curve introduced in section 1 (for a constant markup), while the second is the effect on output of deviations of markups from their steady state value, and hence represents a movement *along* the real marginal cost schedule. We then use this decomposition to investigate the extent to which changes in y are attributable to either term. Because there is no reason to suppose that changes in markups are independent of shifts in the real marginal cost curve, there is more than one way in which this question can be posed. First, one could ask how much of the fluctuations in aggregate activity can be attributed to the fact that markups vary, *i.e.*, would not occur if technology and labor supply varied to the same extent but markups were constant (as would, for example, be true under perfect competition). Alternatively, one might ask how much of these fluctuations are due to markup variations that are not caused by shifts in the real marginal cost schedule, and thus cannot be attributed to shifts in technology or labor supply, either directly or indirectly (through the effects of such shocks on markups).

The first way of posing the question is obviously the one that will attribute the greatest importance to markup variations. On the other hand, the second question is of particular interest, since, as we argued in section 1, we cannot attribute much importance to "aggregate demand" shocks as sources of business fluctuations, unless there is a significant component of output variation at business-cycle frequencies that can be attributed to markup variations in the more restrictive sense.

Mere measurement of the extent to which markup variations are correlated with the cycle – the focus of our discussion in section 2, and the focus of most of the existing literature – does not provide very direct evidence on either question. If we pose the first question, it is obviously necessary that significant markup variations exist, if they are to be responsible for significant variation in economic activity. But the relevant sense in which markup variations must be large is in terms of the size of variations in output that they imply. The size of the correlation of markup variations with output is thus of no direct relevance for this question. Moreover, markup variations could remain important for aggregate activity in this first sense even if markups were procyclical as a result of increasing whenever real marginal costs decline. In this case, markup variations would dampen the effects of shifts in real marginal costs.

If, instead, we ask about the extent to which markup variations contribute to output movements that are independent of changes in real marginal cost, the correlation of markups with output plays a more important role. The reason is that these orthogonal markup fluctuations lead output and markups to move in opposite directions and thus induce a negative correlation between output and markups. However, markups could be very important even without a perfect inverse correlation since, as we show below, the dynamic relationship between markup variations and the employment and output variations that they induce is fairly complex in the presence of adjustment costs. Furthermore, even neglecting this, a strong negative correlation between markups and activity would be neither necessary nor sufficient to establish the hypothesis that orthogonal movements in markups contribute a great deal to output fluctuations. The negative correlation might exist even though the business cycle

is mainly caused by technology shocks, if those shocks induce countercyclical markup variations that further amplify their effects upon output and employment. And the negative correlation might be weak or non-existent even though shocks other than changes in real marginal cost are important, if some significant part of aggregate fluctuations is nonetheless due to these cost shocks, and these shocks induce procyclical markup variations (that damp, but do not entirely eliminate or reverse, their effects upon output).

In this section, we try to settle these questions by carrying out decompositions of the sort specified in (3.17) and analyzing the extent to which y^*, $\hat{y}^\mu$ and the part of $\hat{y}^\mu$ that is orthogonal to y^* contribute to movements in y. We do this for two different measurements of $\hat{\mu}$, which imply different movements in $\hat{y}^\mu$. The first measurement of μ we consider is based on (2.9) while the second is based on the existence of a cost of changing the level of hours worked. Because of space constraints, we are able to give only a cursory and illustrative analysis of these two cases.

We start with the case where markups are given by (2.9), for which we gave several interpretation above. To compute how much output rises when markups fall, we must make an assumption about the extent to which workers demand a higher wage when output rises. We thus assume that, in response to changes in markups, wages are given by

$$\hat{w}_t = \eta_W \hat{H}_t. \tag{3.18}$$

Thus, η_W represents the slope of the labor supply curve along which the economy moves when markups change. Obviously, this simple static representation is just a simplification. We again let η_H represent the elasticity of output with respect to hours when hours are being changed by markup movements. Using (3.18) in (2.9) together with the assumption that changes in output induced by markup changes equal η_H times $\hat{H}$, it follows that

$$\hat{\mu} = -\left(\frac{1-b-\eta_H(1-a)}{\eta_H} + \frac{\eta_W}{\eta_H}\right)\hat{y}^\mu \tag{3.19}$$

where the term in parenthesis is positive because η_H is smaller than one and a and b are nonpositive. This formula allows us to compute $\hat{y}^\mu$ once we have measured $\hat{\mu}$ as above.

In other words, it allows us to go from the measurement of markups to the measurement of output movements implied by markups. Once we have obtained $\hat{y}^{\mu}$ in this manner, we subtract this from y to obtain y^*, as required by (3.17). To do all this, we need three parameters, namely a and b (to construct the markup) and the expression in parenthesis in (3.19). Our illustrative calculation is based on setting a equal to zero, b equal to -.4 (which we saw guarantees that the markup is quite countercyclical) and setting the expression in parenthesis equal to 1/.7. Given these values for a and b, this last parameter can be rationalized by supposing that $\eta_H = .7$ and $\eta_W = .3$. This elasticity of labor supply is broadly consistent with the estimates of Solon, Barsky and Parker (1994).

If we use these parameters and compute $\hat{\mu}$ in the way that we did in Section 2, however, the variance of $\hat{\mu}$ and, in particular, the movements in $\hat{\mu}$ that are orthogonal to movements in $\hat{y}$ are rather large. These orthogonal movements in $\hat{y}^{\mu}$ must then be matched by equal and opposite movements in y^*. One interpretation of this is that shifts in the marginal cost curve would lead to much larger output swings than those we actually observe if it weren't for procyclical markup movements that dampen these shifts. Another interpretation is that there are large errors in the measurement of the wage that lead the labor share to be measured with error. These random movements in the labor share then lead to offsetting movements in the two terms of (3.17), $\hat{y}^{\mu}$ and y^*.

To deal with this possibility, we modify the analysis somewhat. Instead of using actual wages in computing $\hat{\mu}$, we use the projection of the ratio of per capita compensation to per capita output, $(w-y)$, onto the cyclical variables that we used in Rotemberg and Woodford (1996a). In other words, we make use of the regression equation

$$w_t - y_t = \Phi_W Z_t \tag{3.20}$$

where Z_t now represents the current and lagged values of the change in private value added, the ratio of nondurables and services consumption to output, and detrended hours worked in the private sector. To obtain the ratio of per capita compensation to per capita output that we use in (3.20) we divided the labor share by the deviation of hours from their linear

trend. Since this same deviation of hours is an element of the Z_t vector, we would have obtained the same results if we had simply projected the labor share itself. For this included level of hours (and output) to be comparable to the labor share we use to construct $(w-y)$, this labor share must refer to the private sector as a whole. We thus use only this particular labor share in this section. Because of the possibility that this labor share does not follow a single stationary process throughout our sample, we estimated (3.20) only over the sample 1969:1 to 1993:1.

Equation (3.20) allows us to express $(w-y)$ as a linear function of Z. Given that a is zero, the only other determinant of the markup in (2.9) is the level of hours $\hat{H}$, which is also an element of Z. Thus, our estimate of $\hat{\mu}$ is now a linear function of Z. Equation (3.19) then implies that $\hat{y}_t^{\mu}$ is a linear function of Z_t as well. It is not the case, however, that y_t^* is a linear function of Z_t. The reason for this is that Z includes only stationary variables and therefore does not include y. On other hand, the change in private value added, Δy, is an element of Z. This means that, armed with the stochastic process for Z that we estimated in Rotemberg and Woodford (1996a),

$$Z_t = AZ_{t-1} + \epsilon_t \tag{3.21}$$

we can construct the innovations in $\hat{y}^{\mu}$ and in y^*. These are linear functions of the vector ϵ_t which, given (3.21) equals $(Z_t - AZ_{t-1})$ so that these innovations depend only on the history of the Z's. Similarly, the vector $(Z_t - AZ_{t-1})$ together with the matrix A in (3.21) determines by how the expectation of future values of Z is revised at t. This means that we can use (3.21) to write down the revisions at t in the expectations of y_{t+k}, $\hat{y}_{t+k}^{\mu}$ and y_{t+k}^* as linear functions of the history of the Z's. Finally, the variance covariance matrix of the ϵ's (which can be obtained from A and the variance covariance matrix of the Z's) then implies variances and covariances for both the innovations and revisions in the y's, the $\hat{y}^{\mu}$'s and the y^*'s.

Table 3 focuses on some interesting aspects of these induced variances and covariances. Its first row focuses on innovations so that it shows both the variance of the innovation

in y^* and in $\hat{y}^\mu$ as ratios of the innovation variance in y. The subsequent rows focus on revisions at various horizons. The second row, for example, gives the population variances of the revisions at t of y^*_{t+5} and $\hat{y}^\mu_{t+5}$ as ratios to the variance of the revision of y_{t+5}. All these revisions correspond to output changes one year after the effect of the corresponding ϵ_t's is first felt. The next row looks at innovations two years after the innovations first affect output and so on.

We see from table 3 that this measure of the markup has only a very modest effect on one's account of the source of aggregate fluctuations in output. The variances of revisions in y^* are almost equal to the corresponding variances of y for all the horizons we consider. The innovation variance of y^* is actually bigger which implies that innovations in $\hat{y}^\mu$ that are negatively correlated with y^* dampen the effect of these short run movements of y^* on y. The last column in Table 3 looks at the variances of the component of $\hat{y}^\mu$ that is orthogonal to y^*. This variance is equal to the variance of $\hat{y}^\mu$ times $(1-\rho^2)$ where ρ represents the correlation between $\hat{y}^\mu$ and y^* and where this correlation can easily be computed from (3.21). To make the results clearer, we again present the variance of this orthogonal component of $\hat{y}^\mu$ as a fraction of the corresponding variance of y. It is apparent from this column that this orthogonal components explains very little of the variance of y at any of the horizons we consider. Thus, even though this measure of the markup is negatively correlated with our cyclical indicators, it induces movements in output that are much smaller than the actual ones.

Overturning this finding appears to require implausible parameters. To make output more responsive to markup changes requires that the term in parenthesis in (3.19) be smaller. We could achieve this by making η_H smaller or η_W bigger but, given the values that we have chosen, large changes of this sort would be unreasonable. An alternative way of lowering this coefficient is to make b smaller in absolute value. The problem is that, as we saw before, this makes the markup less cyclical. Thus, it does not help in making $\hat{y}^\mu$ track more of the cyclical movements in output. By the same token, setting a equal to a large negative number makes the markup more countercyclical but raises the term in parenthesis in (3.19) thereby

reducing the effect of the markup on $\hat{y}^{\mu}$.

We now turn to the case where adjustment costs imply that deviations of the markup from the steady state are given by (2.15). We follow Sbordone (1996) in that we also let output vary with employee effort and this, as we saw, is consistent with (2.15). Letting a be zero and using (2.14), equation (2.15) can be rewritten as

$$\hat{\mu}_t = \hat{y}_t - \hat{w}_t - \hat{H}_t - c\Big[(\hat{H}_t - \hat{H}_{t-1}) - E_t(\hat{H}_{t+1} - \hat{H}_t)\Big] \tag{3.22}$$

Allowing for variable effort is useful because it relaxes the restriction that the short run output movements induced by markup variations are perfectly correlated with changes in hours. Thus, as in our earlier discussion of her model, we suppose that output is given by $Y = F(K, zeH)$. As a result, we have

$$\hat{y} = \eta_H(\hat{H} + \hat{e}). \tag{3.23}$$

We suppose, as before that the wage bill is given by $H\bar{w}g(e)$ where $\bar{w}$ captures all the determinants of the wage that are external to the firm and g is an increasing function. This leads once again to (2.19) which, once linearized, can be written as

$$\hat{e}_t = \frac{c}{\epsilon_\omega}\Big[(\hat{H}_t - \hat{H}_{t-1}) - E_t(\hat{H}_{t+1} - \hat{H}_t)\Big] \tag{3.24}$$

Finally, we assume that average wages are given by

$$\hat{w}_t = \hat{w}_{ot} + \eta_W \hat{H}_t + \psi\hat{e}_t \tag{3.25}$$

It is important to stress that the parameters η and ψ do no correspond to the elasticities of the average wage paid by an *individual firm* with respect to the *individual firm's* hours and effort level. Rather, they are the elasticities of the economy-wide average wage with respect to aggregate changes in hours and average work effort. Note also that $\hat{w}_{ot}$ is the exogenous component of the wage, *i.e.*, the one that is not affected by changes in markups. Using (3.23) and (3.25) to substitute for $\hat{y}_t$ and $\hat{w}_t$ respectively in (3.22) and then using (3.24) to substitute for $\hat{e}_t$ in the resulting expression we obtain

$$\hat{\mu}_t + \hat{w}_{ot} = (\eta_H - 1 - \eta_W)\hat{H}_t + (\eta_H - \psi - \epsilon_\omega)\frac{c}{\epsilon_\omega}\Big[(\hat{H}_t - \hat{H}_{t-1}) - E_t(\hat{H}_{t+1} - \hat{H}_t)\Big]$$

This difference equation in $\hat{H}$ can be written as

$$E_t \frac{\beta}{L}(1-\tilde{\lambda}_1 L)(1-\tilde{\lambda}_2 L)\hat{H}_t = -\tilde{\zeta}(\hat{\mu}_t - \hat{w}_{ot})$$

where L is the lag operator while $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are the roots of

$$\beta\lambda^2 - [1+\beta+\theta]\lambda + 1 = 0$$

and

$$\theta \equiv \frac{1+\eta_W-\eta_H}{\psi+\epsilon_\omega-\eta_H}\frac{\epsilon_\omega}{c} \qquad \tilde{\zeta} \equiv \frac{1}{\psi+\epsilon_\omega-\eta_H}\frac{\epsilon_\omega}{c}$$

Noting that $\tilde{\lambda}_1\beta$ is equal to $1/\tilde{\lambda}_2$ and letting $\tilde{\lambda}$ be the smaller root (which is also smaller than one as long as $1+\eta_W > \eta_H$ and $\psi+\epsilon_\omega > 0$), the solution of this difference equation is

$$\hat{H}_t = -\tilde{\zeta}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\tilde{\lambda}^k(\beta\tilde{\lambda})^j E_{t-k}[\hat{\mu}_{t+j-k} - \hat{w}_{ot+j-k}] \tag{3.26}$$

The deviations of hours from trend that are due to changes in markups, $\hat{H}_t^\mu$, can then be obtained by simply ignoring the movements in $\hat{w}_o$ in (3.26). We can then compute the deviations of output from trend that are due to markup variations, $\hat{y}^\mu$ by combining (3.23) and (3.24) to yield

$$\hat{y}_t^\mu = \eta_H\Big(\hat{H}_t^\mu + \frac{c}{\epsilon_\omega}\Big[(\hat{H}_t^\mu - \hat{H}_{t-1}^\mu) - E_t(\hat{H}_{t+1}^\mu - \hat{H}_t^\mu)\Big]\Big) \tag{3.27}$$

This implies that, as before, $\hat{y}^\mu$ is a linear function of current and past values Z_t. To see this, note first that (3.22) implies that we can write $\hat{\mu}_t$ as a function of Z_t. The reason for this is that $(w-y)$ is a function of Z_t, $\hat{H}_t$ is part of Z_t and, as a result of (3.21), $E_t\hat{H}_{t+1}$ is the corresponding element of AZ_t. Therefore, using (3.21) once again, the expectation at t of future values of $\hat{\mu}$ must be a function of Z_t. Past expectations of markups which were, at that point, in the future are therefore functions of past Z's. The result is that we can use (3.26) to write $\hat{H}_t^\mu$ as a function of the history of the Z's.[67] Finally, we use (3.27) to write the component of output that is due to markup changes as a function of the Z's.

We require several parameter values to carry out this calculation. First, we set c equal to 8 to calculate $\hat{\mu}_t$ in (3.22). To compute the connection between $\hat{y}^\mu$ and the Z's we need

three more parameters. It is apparent from (3.26) and (3.27) that this calculation is possible if one knows $\tilde{\lambda}$, $\tilde{\zeta}$ and ϵ_ω in addition to c (which is needed to compute markups anyway). For illustrative purposes, we set these three parameters equal to .79, .13 and 3 respectively. The parameters $\tilde{\lambda}$ and $\tilde{\zeta}$ are not as easy to interpret as the underlying economic parameters we have used to develop the model. In addition to c and ϵ_ω these include η_H, η_W, ψ. Because the number of these parameters is larger than the number of parameters we need to compute $\hat{y}^\mu$, there is a one dimensional set of these economically meaningful parameters that rationalizes our construction of $\hat{y}^\mu$. In particular, while this construction is consistent with supposing that η_H, η_W, ψ equal .7, .25 and .1 respectively it is also consistent with different values for these parameters.[68]

We use our knowledge of the relationship between $\hat{y}^\mu$ and the Z's for two purposes. As before, we compute the variances of the innovations and revisions in $\hat{y}^\mu$ as well as of y^*. Second, we look at the resulting sample paths of $\hat{y}^\mu$ and y^*. The second part of Table 3 contains the variances, which correspond to the ones we computed before. The results are quite different, however. In particular, the variance of the component of $\hat{y}^\mu$ that is orthogonal to y^* now accounts for about 90% of the variance of the revisions in output growth over the next two years. Thus, independent markup movements are very important in explaining output fluctuations over "cyclical" horizons. Moreover, if one takes the view that the movements of y that are genuinely attributable to y^* are those which are not due to the component of $\hat{y}^\mu$ that is orthogonal to y^*, the movements in y^* account for only about 10% of the movements in y. Movements in y^* have essentially no cyclical consequences. It is not that the revisions in the expectations of y^* are constant. Rather, upwards revisions in y^* over medium term horizons are matched by increases in markups that essentially eliminate the effect of these revisions on y. This cannot be true over long horizons since the markup is assumed to be stationary so that $\hat{y}^\mu$ is stationary as well. Thus, changes in y^* that are predicted 20 years in advance account for about 80% of the revisions in output that are predicted 20 years in advance.

An analysis of the sample path of $\hat{y}^\mu$ (and the corresponding path of y^*) delivers similar

results. Such sample paths can be constructed since $\hat{y}^{\mu}$ depends on the Z's which are observable. Admittedly, (3.26) requires that the entire history of Z's be used. Data limitations thus force us to truncate k at 18 as explained in footnote 64. The result is that $\hat{y}^{\mu}$ depends on 18 lags of Z. To make sure that the lagged expectations of markups which enter (3.26) are computed within the period where the labor share remains a constant stationary function Z_t, we construct this sample path starting in 1973:2. The resulting values of y^* and the log of output y are plotted in Figure 3. It is apparent from this figure that the episodes that are usually regarded as recessions (which show up in the Figure as periods where y is relatively low) are not reflected in movements of y^*. Figure 4 plots instead $\hat{y}^{\mu}$ against the predicted declines of output over the next 12 quarters. These series are nearly identical so that, according to this measure of the markup, almost all cyclical movements in output since 1973 are attributable to markup variations. This second measure of markups is thus much more successful in accounting for cyclical output movements. This result is probably partly due to the fact that this method of estimating $\hat{y}^{\mu}$ recognizes the possibility that, in booms, output expands more than is suggested by the labor input as a result of increases in effort.[69]

4 Models of Variable Markups

We now briefly review theoretical models of markup variation. We give particular attention to models in which markups vary endogenously, and thus affect the way the economy responds to shocks. The shocks of interest include both disturbances that shift the marginal cost schedule and other sorts of shocks, where these other shocks would not have any effect on equilibrium output in the absence of an effect upon equilibrium markups.

Before reviewing specific models, it is perhaps worth commenting upon the kind of theoretical relations between markups and other variables that are of interest to us. It is important to note that an explanation for countercyclical markups need *not* depend upon a theory that predicts that desired markups should be a decreasing function of the level of economic activity. If the real marginal cost schedule $c(Y)$ is upward-sloping, then *any* source of variations in the markup that is independent of variations in the marginal cost schedule

itself will result in inverse variations in the level of output, and so a negative correlation between the markup and economic activity. Thus theories of why markups should vary as functions of interest rates or inflation (rather than of the current level of economic activity) might well be successful explanations of the cyclical correlations discussed in section 2. In fact, a theory according to which the markup should be a function of the level of economic activity is, in some respects, the least interesting kind of theory of endogenous markup variation. This is because substitution of $\mu = \mu(Y)$ into (1.1) still gives no reason for equilibrium output Y to vary, in the absence of shifts in the marginal cost schedule. (Such a theory, with μ a decreasing function of Y, could however serve to amplify the output effects of shifts in that schedule.)

Care is also required in relating theories of pricing by a particular firm or industry, as a function of conditions specific to that firm or industry, to their implications for aggregate output determination. For example, a theory according to which a firm's desired markup is an increasing function of its *relative* output, $\mu^i = \mu(y^i/Y)$ with $\mu' > 0$, might be considered a theory of "procyclical markups". But in a symmetric equilibrium, in which all firms price according to this rule, relative prices and outputs never vary, and there will be no cyclical markup variation at all. If instead (as discussed in section 4.3 below) not all firms continuously adjust their prices, the fact that adjusting firms determine their desired markup in this way can reduce the speed of overall price adjustment; and this increase in price stickiness can increase the size of the *countercyclical* markup variations caused by disturbances such as changes in monetary policy.

The models we look at fall into two broad categories. In the first class are models where firms are unable to charge the price (markup) that they would like to charge because prices are sticky in nominal terms. Monetary shocks are then prime sources of discrepancies between the prices firms charge and the prices they would like to charge. This leads to changes in markups that change output even if desired markups do not change. In the second class of models, real factors determine variations in desired markups, even in the case of complete price flexibility. Finally, we briefly consider interactions between these two types

of mechanisms.

4.1 Sticky Prices

We do not provide a thorough survey of sticky price models since that is taken up in Taylor (this volume). Rather, our aim is threefold. First, we want to show how price stickiness implies markup variations, and so may explain some of the findings summarized in our previous sections. Second, we want to argue that markup variations are the crucial link through which models with sticky prices lead to output variations as a result of monetary disturbances. In particular, such models imply a link between inflation and markups which is much more robust than the much-discussed link between inflation and output. Thus viewing these models as models of endogenous markup variation may help both in understanding their consequences and in measuring the empirical significance of the mechanisms they incorporate. Finally, we discuss why sticky prices alone do not suffice to explain all of the evidence, so that other reasons for countercyclical markups also seem to be needed.

It might seem at first peculiar to consider output variations as being determined by markup variations in a model where prices are sticky. For it might be supposed that if prices are rigid, output is simply equal to the quantity demanded at the predetermined prices, so that aggregate demand determines output directly. However, this is true only in a model where prices are absolutely fixed. It is more reasonable to suppose that some prices adjust, even over the time periods relevant for business cycle analysis. The issue then becomes the extent to which prices and output adjust, and, as we shall see, this is usefully understood in terms of the determinants of markup variation.

We illustrate the nature of markup variations in sticky-price models by presenting a simple but canonical example, which represents a discrete-time variant of the model of staggered pricing of Calvo (1983), the implications of which are the same (up to our log-linear approximation) as those of the Rotemberg (1982) model of convex costs of price adjustment. First, we introduce a price-setting decision by assuming monopolistic competition among a large number of suppliers of differentiated goods. Each firm i faces a downward-sloping demand

curve for its product of the form

$$Y_t^i = D\Big(\frac{P_t^i}{P_t}\Big)Y_t \tag{4.1}$$

where P_t^i is the price of firm i at time t, P_t is an aggregate price index, Y_t is an index of aggregate sales at t, and D is a decreasing function. We suppose that each firm faces the same level of (nominal) marginal costs C_t in a given period.[70] Then neglecting fixed costs, profits of firm i at time t are given by

$$\Pi_t^i = (P_t^i - C_t)D\Big(\frac{P_t^i}{P_t}\Big)Y_t.$$

Following Calvo, we assume that in each period t, a fraction $(1-\alpha)$ of firms are able to change their prices while the rest must keep their prices constant. A firm that changes its price chooses it in order to maximize

$$E_t \sum_{j=0}^{\infty} \alpha^j R_{t,t+j} \frac{\Pi_{t+j}^i}{P_{t+j}},$$

where $R_{t,t+j}$ is the stochastic discount factor for computing the present values at t of a random level of real income at date $t+j$. (The factor α^j represents the probability that this price will still apply j periods later.) Denoting by X_t the new price chosen at date t by any firms that choose then, the first-order condition for this optimization problem is

$$E_t \sum_{j=0}^{\infty} \alpha^j R_{t,t+j} \frac{Y_{t+j}}{P_{t+j}} D'\left(\frac{X_t}{P_{t+j}}\right) \frac{X_t}{P_{t+j}} \left[1 - \frac{1}{\epsilon_D(X_t/P_{t+j})} - \frac{C_{t+j}}{X_t}\right] = 0, \tag{4.2}$$

where $\epsilon_D(x) \equiv -xD'(x)/D(x)$ is the elasticity of the demand curve (4.1).

For now, we further simplify by assuming that the elasticity of demand is a positive constant (as would result from the kind of preferences over differentiated goods assumed by Dixit and Stiglitz, 1977). This means that a firm's desired markup, in the case of flexible prices, would be a constant, $\mu^* = \epsilon_D/(\epsilon_D - 1)$. In this way we restrict attention to markup variations due *purely* to delays in price adjustment. It is useful to take a log-linear approximation of the first-order condition (4.2) around a steady state in which all prices are constant over time and equal to one another, marginal cost is similarly constant, and the constant ratio of price to marginal cost equals μ^*. Letting $\hat{x}_t$, $\hat{\pi}_t$, and $\hat{c}_t$ denote percentage deviations

of the variables X_t/P_t, P_t/P_{t-1}, and C_t/P_t respectively from their steady-state values, (4.2) becomes

$$E_t \sum_{j=0}^{\infty} (\alpha\beta)^j \left\{ [\hat{x}_t - \sum_{k=1}^{j} \hat{\pi}_{t+k}] - \hat{c}_{t+j} \right\} = 0, \tag{4.3}$$

where $\beta < 1$ is the steady-state discount factor. Here the factor

$$\hat{x}_t - \sum_{k=1}^{j} \hat{\pi}_{t+k}$$

represents the relative price in period $t+j$ of the firm that chooses new price X_t in period t, and so (4.3) says, essentially, that the firm's price is expected to be proportional to its marginal cost of production *on average*, over the time that the price chosen at date t applies. This equation can be solved for the relative price $\hat{x}_t$ of firms that have just changed their price, as a function of expected future inflation and real marginal costs. The resulting relation can be quasi-differenced to yield

$$\hat{x}_t = \alpha\beta E_t \hat{\pi}_{t+1} + (1 - \alpha\beta)\hat{c}_t + \alpha\beta E_t \hat{x}_{t+1}. \tag{4.4}$$

We suppose that the price index P_t is a symmetric homogeneous degree one function of the prices of the individual goods. Then near the steady state, it can be approximated to first order by the geometric average of the prices. Since each period a fraction α of the prices remain unchanged, while the others all change to the common value X_t, the rate of increase of the index satisfies

$$\hat{\pi}_t = (\frac{1-\alpha}{\alpha})\hat{x}_t$$

in our log-linear approximation. Substituting this into (4.4), we obtain

$$\hat{\pi}_t = \beta E_t \hat{\pi}_{t+1} - \kappa \hat{\mu}_t, \tag{4.5}$$

where $\kappa \equiv (1-\alpha\beta)(1-\alpha)/\alpha$ and $\hat{\mu}_t = -\hat{c}_t$ denotes the percentage deviation of the average markup $\mu_t \equiv P_t/C_t$ from its steady-state value of μ^*.

This equation relates the average markup at any date to current and expected future inflation. To obtain the behavior of equilibrium output, one must use equation (1.1) along

with this. If we log-linearize the real marginal cost schedule as $\hat{c}_t = \eta_c \hat{Y}_t$, where $\hat{Y}_t$ denotes the percentage deviation of output from trend, then (1.1) implies $-\hat{\mu}_t = \eta_c \hat{Y}_t$. Substitution of this into (4.5) then yields an aggregate supply relation of the form

$$\hat{\pi}_t = \theta \hat{Y}_t + \beta E_t \hat{\pi}_{t+1}, \tag{4.6}$$

where $\theta \equiv \kappa \eta_c$. This equation specifies an upward-sloping relation between inflation and output variations, for any given level of expected inflation. Roberts (1995), who specifies $\beta = 1$, calls this "the New Keynesian Phillips Curve," and provides econometric evidence that (when extended to allow for stochastic shifts in the real marginal cost schedule) U.S. output and inflation data are consistent with a relation of this kind.

Combined with a specification of the evolution of aggregate nominal spending (which is often taken, as for example in Rotemberg (1996), to be an exogenous process determined by monetary policy), equation (4.6) allows us to solve for equilibrium fluctuations in output. Because variations in inflation must be associated with deviations of output from trend, monetary policy disturbances affect equilibrium output. It will be observed that the output fluctuations in response to such shocks are associated with countercyclical variations in the average markup.

The endogenous markup variations affect the predicted response of output to other shocks as well. For example, technology shocks may be considered, by allowing for a stochastic shift term in the real marginal cost schedule. Such shocks may be associated with procyclical markup variations: a technological improvement lowers marginal cost, but because prices do not fall immediately in proportion to the decline in costs, markups rise, while (because prices do fall some) output increases. This is consistent with (4.5) and (4.6) if prices fall faster immediately than they are expected to in the future. In such a case, the markup variation damp the output effects of the technology shocks relative to what would happen under perfect competition; as a result, input demand may actually decline in response to a favorable technology shock.[71]

We have seen that a sticky-price model of this kind involves endogenous variation in the

average markup. But is it useful to think of the endogenous markup variations as central to the way that nominal variables have real effects in this model? We believe that there are several advantages to viewing the model in this way (in addition, of course, to the fact that it helps one in relating the predictions of the sticky-price model to the kinds of facts discussed in sections 2 and 3). First, if one is willing (as seems reasonable) to abstract from the effects of monetary frictions upon the relations (labor supply, labor demand, and so on) that underlie the real marginal cost schedule, then the effects of monetary policy upon the determination of real variables may be reduced entirely to its effects upon the average markup. A general equilibrium model of the effects of monetary policy may then be usefully decomposed into three distinct parts, each derived from largely separate microeconomic foundations: (i) a theory of equilibrium output determination *given* the markup, essentially a more elaborate version of equation (1.1); (ii) an equation relating the markup to inflation variations, equation (4.5); and (iii) a theory of how monetary policy affects nominal aggregate demand. An advantage of viewing the structure of one's macroeconomic model this way is that part (i) of the model involves no specifically monetary elements, and may (except for the allowance for a time-varying markup) be identical to the equations of a real business cycle model, while part (iii) does not involve the specification of aggregate supply relations, and so may be directly adapted from conventional Keynesian or monetarist models of the effects of monetary policy on aggregate demand. The theory of endogenous markup variation thus provides the crucial link that allows the concerns of real business cycle models and conventional monetary models to be synthesized.[72]

Second, understanding the markup variations that are associated with variations in real activity in a sticky-price model is important to understanding when and how those fluctuations in activity are inefficient, since the markup directly measures the extent to which a condition for efficient resource allocation fails to hold. This perspective can be a source of important insights into the welfare losses from price-level instability and the nature of optimal monetary policy.

And third, recognizing that (4.5) is a more fundamental prediction of the model of price-

setting than is a relation such as (4.6), which also depends upon one's specification of wage-setting behavior and the like, may allow more accurate empirical estimation of the speed of aggregate price adjustment. Sbordone (1997) tests the accuracy of (4.5) as a model of aggregate price dynamics in the U.S. by first estimating the evolution of unit labor cost (assumed to be proportional to marginal cost) using a VAR. Using this evolution of unit labor costs, she then computes the equilibrium path of the price index implied by (4.5). She finds that this simple model accounts quite well for the evolution of the private GDP deflator in the U.S., at the quarterly frequency, over the period 1960-1997. In the case of her best-fitting value for α,[73] the variance of the discrepancy between the actual price series and the one that would be predicted on the basis of the unit labor cost process is reduced to only 12% of what it would be in the absence of price rigidity,[74] while the variance of the discrepancy between the actual and predicted inflation series is reduced to only 4% of what it would be according to the flexible-price (constant-markup) model. It is especially striking that the model fits this well without any need for complications such as stochastic disturbances to the pricing equation; this suggests that (4.5) is indeed more descriptive of the data than the aggregate supply relation (4.6). This would suggest that the stochastic disturbances to this aggregate supply relation, which require Roberts (1995) to add additional terms to his estimated equation and to estimate it using instrumental variables, represent mainly disturbances to the real marginal cost schedule, rather than disturbances to the pricing relation (4.5).

Despite the impressive success of this simple model as an explanation of much of the cyclical variation in prices relative to labor costs, there is some reason to doubt that this model of markup variation is completely adequate. In particular, the implication that the output effect of supply shocks is muted in sticky price models is problematic given that, as suggested by Hamilton (1983) economic activity has tended to fall in the aftermath of pre-1986 oil price increases. If the principal effect of oil price increases is to increase marginal costs, then a sticky price model (by implying that prices should rise less than the increase in marginal costs, so that markups fall) will imply even less of a contraction of equilibrium

output than one should expect in the case of a flexible-price model. However, the size of the observed contractionary effects of oil price shocks on the U.S. economy is already rather larger than makes sense under competitive pricing, owing the relatively small share of energy costs in total marginal costs of production. For this reason, Rotemberg and Woodford (1996b) propose that oil price increases lead to *increases* in desired markups. With this motivation, we turn to a brief review of models where desired markups vary.

4.2 Variations in Desired Markups

For simplicity, in this section we assume completely flexible prices. We also simplify by making all firms fully symmetric so that, in equilibrium, they all charge the same price. A number of types of theories of this kind have been considered in the literature.

Varying Elasticity of Demand.

Probably the simplest and most familiar model of desired markup variations attributes them to changes in the elasticity of demand faced by the representative firm. There are two important ways in which one might allow for variations in the elasticity of demand at a symmetric equilibrium where all relative prices are equal to one.

The first is to suppose that the utility and/or production functions that define buyers' preferences over differentiated goods are not homothetic, so that changes in aggregate purchases Y_t change the elasticity of demand. This is not an entirely satisfactory assumption, however, because it is unappealing to assume that growth should lead to secular changes in the elasticity of demand and in markups. One may, however, avoid this implication by complicating the model, for example by assuming that growth is associated with an increase in the number of differentiated goods rather than any secular increase in the scale of production of any individual goods

A more appealing way of obtaining changes in this elasticity is to follow Gali (1994) and Heijdra (1995) and assume that there are several different kinds of purchasers.[75] Each of these purchases all of the goods that are produced, but the different types each have a different preferences over differentiated goods. Suppose, for example, that two groups 1 and

2 each care only about the amount they obtain of a composite good defined by a symmetric, homogeneous degree one aggregate of all goods purchases, but that the aggregator functions H_1 and H_2 are different for the two groups. Then the demand for good i by group j can be written as $Y_{j,t}D_j(P_t^i/P_{j,t})$ where $Y_{j,t}$ is the quantity purchased by group j of its composite good, and $P_{j,t}$ is the price of that composite good (a homogeneous degree one index of the individual goods prices). Total demand for good i is then

$$Y_t^i = D_1\Big(\frac{P_t^i}{P_{1,t}}\Big)Y_{1,t} + D_2\Big(\frac{P_t^i}{P_{2,t}}\Big)Y_{2,t}, \tag{4.7}$$

where $D_j(1) = 1$ for both groups. At a symmetric equilibrium, all prices are the same and the amount purchased of each good is the same, so that $Y_{j,t}$ is simply the amount purchased of each good by group j. The elasticity of demand at a symmetric equilibrium is then found to be

$$Z_tD_1'(1) + (1 - Z_t)D_2'(1) \qquad \text{where} \qquad Z_t = \frac{Y_{1,t}}{Y_{1,t} + Y_{2,t}}.$$

Therefore, an increase in the share of group 1 purchases in total purchases makes the overall elasticity of demand more similar to D_1', the elasticity of the demand by group 1. An important feature of business cycles is that, as noted in Campbell (1987) and Cochrane and Sbordone (1988), the ratio of consumption to GDP is high in recessions and low in booms. Exactly the converse behavior applies to the ratio of investment to GDP. Thus, as Gali (1992) points out, the assumption that firms have more elastic demands than consumers can provide one explanation for countercyclical markups. Moreover, an exogenous increase in the fraction of output demanded for investment purposes would increase y^μ.

Another variable that varies more cyclically than GDP is the purchase of durables. This has led Bils (1989) and Parker (1996) to argue that increased purchases of durables in booms reduce markups in these periods. This idea is closely related to a proposal of Robinson (1932), who argued that people who purchased durables in downturns were predominantly replacing durables that had ceased functioning and that, as a result, demand in downturns was less elastic than demand in booms, which consisted largely of demand by new purchasers.

To ensure that a story of this sort also leads to reduced markups when the government

expands its own purchases of goods and services, as would be needed in order to account for the expansionary effects of government purchases other than through an effect on labor supply, one must assume that the government has a relatively elastic demand.[76] The main disadvantage of this general type of explanation is that aggregate demand, as such, has no direct role in lowering markups and thereby increasing output. Rather, it is the *composition* of demand that affects aggregate output; increases in aggregate demand only raise output if they happen to shift demand towards sectors with more elastic demand. This means that at least some kinds of disturbances that increase some important component of current spending must be contractionary rather than expansionary (*e.g.*, an increase in consumer demand, in Gali's model). It is hard to think of empirical support for this kind of prediction.

Customer Markets.

An alternative class of models, that gives variations in aggregate demand a more direct role, is *intertemporal* models of markup variation, in which what matters is not the composition of demand at present, but rather the how current sales compare to expected future sales. Probably the best-known model of this type is the "customer market" model of Phelps and Winter (1970).

The customer market model is a model of monopolistic competition, in that each firm maximizes profits with respect to its own price (markup) taking the price (markup) of all other firms as given. But it differs from the standard model of monopolistic competition (*e.g.*, the model of Dixit and Stiglitz (1977)) in introducing a dynamic element into the response of demand to prices. A firm that lowers its current price not only sells more to its existing customers, but also expands its customer base. Having a larger customer base leads future sales to be higher at whatever price is charged then. One simple formulation that captures this idea involves writing the demand for firm i at time t as

$$Y_t^i = Y_t \eta\Big(\frac{\mu_t^i}{\mu_t}\Big) m_t^i \qquad \eta' < 0, \quad \eta(1) = 1, \tag{4.8}$$

where μ_t^i is the markup of price over marginal cost implicit in the price charged by firm i at time t, and the "market share" m_t^i is the fraction of average demand Y_t that goes to firm

i if it charges the same price as all other firms. The market share depends on past pricing behavior according to the rule

$$m^i_{t+1} = g\Big(\frac{\mu^i_t}{\mu_t}\Big) m^i_t \qquad g' < 0, \quad g(1) = 1, \tag{4.9}$$

so that a temporary reduction in price raises firm i's market share permanently.

Equations (4.8) and (4.9) are intended to capture the idea that customers have switching costs, in a manner analogous to the models of Gottfries (1986), Klemperer (1987), and Farrell and Shapiro (1988).[77] A reduction in price attracts new customers who are then reluctant to change firms for fear of having to pay these switching costs. One obvious implication of (4.8) and (4.9) is that the long run elasticity of demand, *i.e.*, the response of eventual demand to a permanent increase in price, is larger than the short run elasticity of demand. In our case, a firm that charges a higher price than its competitors eventually loses all its customers, though this is not essential for our analysis.

The firm's expected present discounted value of profits from period t onward is thus

$$E_t \sum_{j=0}^{\infty} R_{t,t+j} \Big(\frac{\mu^i_{t+j} - 1}{\mu_{t+j}}\Big) Y_{t+j} \eta\Big(\frac{\mu^i_{t+j}}{\mu_{t+j}}\Big) m^i_t \prod_{z=0}^{j-1} g\Big(\frac{\mu^i_{t+z}}{\mu_{t+z}}\Big).$$

Firm i chooses μ^i_t to maximize this expression, taking as given the stochastic processes $\{\mu_t\}$ and $\{Y_t\}$ that define aggregate demand conditions. Therefore

$$\begin{aligned} Y_t \eta\Big(\frac{\mu^i_t}{\mu_t}\Big) + Y_t \eta'\Big(\frac{\mu^i_t}{\mu_t}\Big)\Big[\frac{\mu^i_t - 1}{\mu_t}\Big] \quad + \\ g'\Big(\frac{\mu^i_t}{\mu_t}\Big) E_t \sum_{j=1}^{\infty} R_{t,t+j} \quad Y_{t+j} \quad \Big[\frac{\mu^i_{t+j} - 1}{\mu_{t+j}}\Big] \eta\Big(\frac{\mu^i_{t+j}}{\mu_{t+j}}\Big) \prod_{z=1}^{j-1} g\Big(\frac{\mu^i_{t+z}}{\mu_{t+z}}\Big) = 0 \end{aligned} \tag{4.10}$$

where subscripts denote partial derivatives. At a symmetric equilibrium where all firms charge the same price, each has a share $m^i_t = 1$, and g equals one in all periods. So the expectation term in (4.10) is equal to the common present discounted value of future profits, which we denote by X_t. Solving (4.10) for the markup, we obtain

$$\mu_t = \mu(X_t/Y_t) \equiv \frac{\eta'(1)}{1 + \eta'(1) + g'(1)\frac{X_t}{Y_t}}. \tag{4.11}$$

Because $\eta'(1)$ and $g'(1)$ are both negative, the derivative of μ with respect to X/Y is negative.[78] An increase in X/Y means that profits from future customers are high relative to profits from current customers so that each firm lowers its price in order to increase its market share. Thus, in this model, expansionary fiscal policy (which raises real interest rates and thus lowers X/Y) raises markups and lowers output.[79] On the other hand, this is a model that can potentially amplify the expansionary effects of loose monetary policy in the presence of sticky prices. The reason is that loose monetary policy lowers real interest rates if prices are rigid and this raises X/Y.[80]

A rather different view of the determinants of markups and output is obtained if the customer market model is combined with the assumption that financial markets are imperfect, as in Greenwald, Stiglitz and Weiss (1984) and Gottfries (1991). With imperfect capital markets, shocks that raise the shadow cost of funds by making it more difficult to borrow (such as reductions in asset values that lower the value of firm's collateral) can lower X/Y and thereby lower output.

Chevalier and Scharfstein (1995, 1996) provide some evidence for this finance-constrained version of the customer market model. Chevalier and Scharfstein (1996) consider pricing by supermarkets and pay particular attention to the prices charged in states hit hard by the oil price decline of 1986. They ask whether, within these oil states, supermarkets that belonged to national chains (and who thus could rely on externally provided cash to some extent) lowered their prices relative to those of local supermarkets, who were presumably more strapped for cash. They find that they do suggesting that national supermarkets were more willing to invest in customers at this point, presumably because they had lower discount rates as a result of their access to cash. Chevalier and Scharfstein (1995) shows more generally that industries with a relatively large fraction of output produced by small firms tend to have more countercyclical markups if one controls for total concentration. The idea is that small firms have less access to capital markets and so should be more strapped for cash in recessions. This induces them to invest less in customers and raise their markups in recessions. The control for concentration creates problems of interpretation because, as discussed further

below, highly concentrated sectors (in which large firms are clearly important) have more countercyclical markups.

Implicit Collusion.

An alternative intertemporal model, where the same variable X/Y again turns out to be the crucial determinant of the equilibrium markup, is the implicit collusion model presented in Rotemberg and Woodford (1992). We consider an economy with many industries, each of which consists of n firms. The n firms in each industry collude implicitly in the sense that there is no enforceable cartel contract, but only an implicit agreement that firms that deviate from the collusive understanding will be punished. On the other hand, the firms in each industry, even when acting in concert, take other industries' prices, the level of aggregate demand, and the level of marginal cost as given. Abusing the language somewhat, we can view industries as monopolistic competitors in the usual sense, while the firms within each industry collude implicitly.

Keeping this distinction in mind, we write the demand for firm i in industry j as

$$Y_t^{ij} = Y_t D^i\Big(\frac{\mu_t^{1j}}{\mu_t}, \ldots, \frac{\mu_t^{nj}}{\mu_t}\Big) \qquad D^i(1, \ldots, 1) = 1. \tag{4.12}$$

The function D^i is symmetric in its first n arguments except the ith, and the functions D^i (for $i = 1, \ldots, n$) are all the same after appropriate permutation of the arguments. The resulting profits of firm i in industry j if all other firms in the economy charge a markup μ_t and it charges a markup μ_t^{ij} are

$$\Pi_t^{ij}(\mu_t, \mu_t^{ij}) = \Big(\frac{\mu_t^{ij} - 1}{\mu_t}\Big) Y_t D^i\Big(\frac{\mu_t^{ij}}{\mu_t}\Big) \tag{4.13}$$

If the firm goes along with the collusive agreement at t and charges the same markup as all other firms, it gets $\Pi_t^{ij}(\mu_t, \mu_t)$ which we denote by $\Pi_t^a(\mu_t)Y_t$. If it deviates and the punishment is as strong as possible, it earns some higher profits at t but it can expect to earn a present value of zero thereafter. In this case, a deviating firm simply maximizes (4.13) with respect to μ_t^{ij} and its resulting profits are $\Pi_t^d(\mu_t)Y_t$. It is easy to show that the difference $\Pi_t^d(\mu_t) - \Pi_t^a(\mu_t)$ is increasing in μ_t. Intuitively, it should be clear that this difference

is zero at the markup that corresponds to the equilibrium where each firm behaves like a monopolistic competitor and takes other firm's prices as given. If firms in the industry charge higher markups, deviating by cutting prices is more attractive. Because this difference is increasing in the markup, a profit maximizing collusive oligopoly which is unable to sustain the monopoly outcome for the industry will agree upon a markup that keeps firms just indifferent between charging the collusive markup and deviating. Such a collusive optimum implies that

$$\Pi_t^d(\mu_t) - \Pi_t^a(\mu_t) = \frac{X_t}{Y_t}. \tag{4.14}$$

This equation can again be solved for an equilibrium markup function of the form $\mu_t = \mu(X_t/Y_t)$.

In Rotemberg and Woodford (1992) we give the conditions under which there exists an equilibrium where (4.14) is binding near a deterministic steady state. Because the left hand side of (4.14) is increasing in the markup μ_t the equilibrium markup is increasing in X/Y. An increase in X, the expected present value of future collusive profits, makes firms want to go along with the collusive price so that this price can be higher. An increase in current output, by contrast, tends to reduce the markup that can be sustained without breaking the collusive agreement. The result is that tight fiscal policy, which raises real interest rates, raises markups and lowers output. Temporary oil price increases also raise X relative to Y and thus also reduce output according to this model.

Rotemberg and Woodford (1991) provide evidence that, if asset price data are used to compute X, markups are not just decreasing in Y but are also increasing in X. This fits well with the finding of Galeotti and Schiantarelli (1995) that markups fall when the expected rate of growth of output is high. Such a high rate of growth raises X since profits are procyclical and this should lead to an increase in markups according to this model.

A striking confirmation that high levels of X raise current markups is provided by Borenstein and Shepard (1996) in their analysis of retail gasoline markets. Their analysis looks at retail gasoline prices in 59 cities over 72 months and takes advantage of the fact that the relationship between expected future demand and current demand varies across cities

because they experience different seasonal cycles. Similarly, there are cross city differences in expected future costs. Borenstein and Shepard show that, consistent with this model, high expected future demand and low expected future costs, both of which raise X, raise current markups. A similar finding, though the evidence in this case is so weak that one cannot reject the hypothesis of no effect, is reported by Ellison (1994). He shows that a railroad cartel operating in the 1880's, the Joint Executive Committee, tended to have low prices when demand was low relative to expected future demand.

The dependence of markups on X leads Bagwell and Staiger (1995) to conclude that this model actually implies procyclical markups. This conclusion follows from identifying booms with periods where the rate of growth of output is high and identifying recessions with periods where the rate of output growth is low. Given that the rate of growth of output is positively serially correlated, periods where output growth is high are actually periods where a crude computation of X/Y (one that only took note of the positive serial correlation of output growth) is high and the conclusion follows. As noted by several authors (see Evans and Reichlin (1994), Rotemberg and Woodford (1996) and the papers cited therein) there are variables other than current output growth that are useful for forecasting future output growth. As Evans and Reichlin (1994) and Rotemberg and Woodford (1996) show, once these other variables are taken into account when computing expected output growth, recessions as defined by the NBER are actually periods where expected output growth is high. Once this is recognized, the model does indeed predict that markups should be high in periods that are generally regarded as recessions.

Because past output growth is nonetheless also somewhat useful in forecasting future output growth, it follows that expected output growth just after business cycle troughs (when output has already started to increase) is higher than expected output growth just before these troughs. Thus, X/Y is higher just after business cycle troughs than just before. The model is thus consistent with some interesting observations made by Baker and Woodward (1994). They compare the price charged by firms some time before an industry trough (the reference month) with the price charged after the trough in the first month in which output

is no smaller than output in the reference month. They report that, for some industries, the latter price is much greater than the former. Moreover, the size of this price increase is larger in more concentrated industries. This suggests that concentrated industries, where this theory is more likely to apply, are ones where the markup is more likely to vary positively with X/Y.

One open question about this model (and the customer market model) is whether they can explain the reduction in inputs that seems to accompany periods of genuine technical progress. What determines which of these two models can explain this fact is whether genuine technical progress raises or lowers X/Y. If the progress raises mainly output in the future, one might expect X to rise relative to Y except that this effect might be offset by an increase in the rate of interest (which reduces the present value X). If X/Y nonetheless rises with technical progress, the implicit collusion would also imply that such shocks tend to raise markups and reduce output relative to what would occur under frictionless perfect competition.

Variable Entry.

A final theoretical reason for markups to vary with cyclical variables is that entry is procyclical. An advantage of this explanation is that it is undoubtedly true that more new firms incorporate in booms, as noted by Chatterjee, Cooper and Ravikumar (1993) for the United States, and documented by Franck (1995) for France. Moreover, as long as profits are procyclical, it makes sense that entry should be procyclical. As we saw in section 3, such procyclical profits are possible even if output fluctuations are entirely due to changes in markups, rather than to shifts in the real marginal cost schedule.

Suppose that, as in Chatterjee, Cooper and Ravikumar (1993) or Portier (1995) firms behave in Cournot fashion so that each industry contains several firms producing perfect substitutes and these firms take the output of all other firms as given when deciding on their own level of output. In this model, the addition of new firms cause markups to fall.[81] The biggest problem with this explanation for countercyclical markups is that technical progress

would lead markups to fall both in the short run and in the long run. As long a technological progress does not increase the fixed cost Φ, such long term progress increases the number of firms and thereby reduces markups.

One way of avoiding this difficulty is to assume that entry simply leads to an increased number of goods being produced by monopolistically competitive firms, as in Devereux, Head and Lapham (1996) or Heijdra (1995). These authors assume that the monopolistically competitive firms produce intermediate goods that are purchased by firms which combine them into final goods by using a Dixit-Stiglitz (1977) aggregator. The result is that increased entry does not change the ratio of price to marginal cost. It does, however, reduce the price of final goods relative to the price of intermediate goods, because final gods can be produced more efficiently when there are more intermediate goods. This reduction in the price of final goods effectively raises real wages and, particularly if it is temporary, leads to an increase in labor supply. Thus, Devereux, Head and Lapham (1996) show that, in their model, an increase in government purchases raises output together with real wages. The increase in output comes about because the increased government purchases make people feel poorer and this promotes labor supply; this results in a shift out of the real marginal cost schedule. The real wage in terms of final goods then rises because of the increase in the number of intermediate goods firms.

4.3 Interactions Between Nominal Rigidities and Desired Markup Variations

Finally, we briefly consider the possibility that markups vary *both* because of delays in price adjustment and because of variations in desired markups, for one or another of the reasons just sketched. This possibility is worth mentioning because interactions between these two mechanisms of markup variation sometimes lead to effects that might not be anticipated from the analysis of either in isolation.

For example, variations in desired markups may amplify the output effects of nominal rigidities, and further slow down the adjustment of prices, even if the corresponding model

of desired markup variation would not imply any interesting effects of monetary policy in the case of flexible prices. To illustrate this, let us consider again the discrete-time Calvo model of section 4.1, but now drop the assumption that the function D has a constant elasticity with respect to the relative price. In this case, log-linearization of (4.2) yields

$$E_t \sum_{j=0}^{\infty} (\alpha\beta)^j \left\{ [\hat{x}_t - \sum_{k=1}^{j} \hat{\pi}_{t+k}] - [\hat{\mu}_{t+j}^{des} + \hat{c}_{t+j}] \right\} = 0, \tag{4.15}$$

as a generalization of (4.3), where $\hat{\mu}_t^{des}$ denotes the percentage deviation of the desired markup

$$\mu_t^{des} \equiv \frac{\epsilon_{D,t}}{\epsilon_{D,t} - 1}$$

from its steady-state value. The elasticity ϵ_D, and hence the desired markup, is a function of the relative price of the given firm i, or equivalently of the firm's relative sales Y^i/Y. Letting the elasticity of the desired markup with respect to relative sales be denoted ξ, we obtain

$$[\hat{x}_t - \sum_{k=1}^{j} \hat{\pi}_{t+k}] - [\hat{\mu}_{t+j}^{des} + \hat{c}_{t+j}] = (1 + \xi\epsilon_D)[\hat{x}_t - \sum_{k=1}^{j} \hat{\pi}_{t+k}] - \hat{c}_{t+j}, \tag{4.16}$$

as a consequence of which (4.15) implies an equation of the same form as (4.3), but with the variable $\hat{c}_t$ replaced by $(1 + \xi\epsilon_D)^{-1}\hat{c}_t$ each period. This in turn allows us to derive again an equation of the form (4.5), except that now

$$\kappa \equiv \frac{1}{1 + \xi\epsilon_D} \frac{(1 - \alpha\beta)(1 - \alpha)}{\alpha}. \tag{4.17}$$

A number of authors have proposed reasons why one might have $\xi > 0$, *i.e.*, an elasticity of demand decreasing in the firm's relative price. Kimball (1995) shows how to construct aggregator functions that lead to arbitrary values of ξ. Thus, this model can rationalize extreme price stickiness even when the fraction of firms that change prices is relatively high. Woglom (1982) and Ball and Romer (1990) suggest that search costs provide an alternative rationale for a positive ξ. The idea is that search costs imply that relatively small price increases lead many customers to depart while small price reductions only attract relatively few customers. A smoothed version of this kinked demand curve gives the variable elasticity just hypothesized.

Equation (4.17) implies that $\xi > 0$ makes κ a smaller positive quantity, for any given assumed average frequency of price changes. This affects the parameters of the markup equation (4.5), and hence the aggregate supply curve (4.6), in the same way as would a larger value of α.[82] In particular, it implies that a given size permanent increase in nominal aggregate demand (due, for example, to a monetary policy shock) will result in both a larger and a more persistent increase in output. Thus allowing for variation in desired markups of this kind can increase the predicted real effects of monetary policy (including the size of the countercyclical markup variations caused by monetary shocks).

To gain some intuition for this result, imagine an increase in aggregate demand which increases marginal cost by increasing the demand for labor by firms whose prices are fixed. A firm which is free to change its prices would thus choose a price above that charged by other firms. If having a price that is relatively high implies that demand is relatively elastic, then such a firm would have a relatively low desired markup and would choose a price that is not far above the one charged by firms with fixed prices. The effect of this is that prices do not rise by as much on impact so that output increases by more. In subsequent periods, the same logic leads those who firms who can change their prices to raise them to only a limited extent. Thus, the effects of the increase in nominal aggregate demand are drawn out.

This occurs despite the fact that the hypothesis of a demand elasticity decreasing in a firm's relative price does not, by itself, provide any reason for monetary policy to have real effects. Indeed, under the hypothesis that all prices are perfectly flexible, it provides no reason for equilibrium markups to vary at all. For with flexible prices, we would expect a symmetric equilibrium in which all firms' prices are always the same, so that the elasticity of demand faced by a firm (and hence its desired markup) would never vary in equilibrium. Thus this hypothesis is much more interesting, both as an explanation of markup variations and as a channel for real effects of shocks other than cost shocks, when combined with the hypothesis of nominal price rigidity than it is on its own.

It is also interesting to note that this hypothesis requires that desired markups be low when the firm's relative price is high, *i.e.*, when its own sales are low relative to those of

its competitors. Thus, its desired markups are *positively* correlated with its own output relative to that of its competitors. At the firm or industry level, one might well observe *procyclical* markups, if one measures the correlation with own output; yet as shown above, the hypothesis is one that can increase the size of the *countercyclical* markup variations at the aggregate level that occur as a result of aggregate demand variations.

Inflation, search and markups are also linked in the work surveyed in Benabou (1992). The idea in this research is that price rigidity in the face of inflation leads to more price dispersion and this price dispersion makes search generally more valuable to consumers. This, in turn, makes demand more elastic for all producers and thus exerts downwards pressure on markups. This theory implies that inflation ought to be generally negatively related to markups. As Benabou (1992) shows, this implication is confirmed in U.S. data on the retail trade sector.

Variations in desired markups that are uniform across goods (rather than depending on firms' relative demands) also interact in interesting ways with nominal rigidity. For example, Kiley (1997) develops a model which combines staggered price-setting with Gali's (1994) assumption of differential demand elasticities for consumption and investment purchases. Monetary expansions then increase investment disproportionately and this temporarily lowers desired markups for all firms. This means that firms that revise their prices do not raise them as much as they otherwise would (given the increase in marginal cost) so that output rises more. This mechanism increases the degree of strategic complementarity among different firms pricing decisions. If other firms raise their prices less, a given change in nominal rates (or in the money supply) has a bigger effect on real rates of interest thereby affecting investment demand more. This, in turn, implies that any given firm wants to raise its price by less. The greater degree of strategic complementarity implies a slower adjustment of the aggregate price level and hence a more persistent effect of the monetary expansion. Thus, while Gali's (1994) model of markup variation does not directly imply that monetary shocks affect output, it increases both the size and the persistence of the output effects of monetary disturbances in the presence of sticky prices.

These illustrations demonstrate that a combination of endogenous variation in desired markups and price stickiness can yield further channels through which disturbances other than cost shocks affect the level of economic activity. This relatively unexplored topic surely deserves further research.

5 Conclusions

The main benefit of allowing for markup variations is that it expands the range of types of disturbances that can affect aggregate economic activity.[83] Without variable markups, output can only increase if real marginal cost falls, for example due to a change in the effective labor supply to firms, or as a result of technological progress. With variable markups, monetary and fiscal shocks can have effects other than those that result from changes in the real wages at which workers are willing to work. In addition, the output effects of certain supply shocks (like variations in the rate of technical progress) may be muted, while other supply shocks (such as oil price increases) can lead to larger output movements.

This rich set of possibilities arises from consideration of a number of different models of variable markups. But it is not clear yet whether there is a single unified model that can make sense of the way all the major macroeconomic shocks affect output. Each of the models we have considered, on its own, seems unable to account for all the facts. In particular, as we have already suggested, the pure sticky price model cannot easily explain the strong effects of oil price increases. The implicit collusion model, on its own, tends to imply that insofar as monetary contractions raise real rates of interest, they should raise rather than reduce output. The customer market model seems to require financial market imperfections to explain the expansionary effects of fiscal stimuli. And it is not clear even then whether it is able to explain the effect of oil and technology shocks on the economy. Thus, the task of constructing a unified model of variable markups that explains the effect of all the shocks we have considered remains to be carried out. Models that allow for interaction between sources of variation in desired markups with additional variation in actual markups due to delays in price adjustment would seem an important area for further study.

Much research remains to be done on the measurement of markup variations as well. First, as we have seen, measurements of markup variations and of the extent to which output fluctuations can be attributed to them depend on the details of the production structure. For example, they are extremely sensitive to the presence of adjustment costs for employment. While the existence of such adjustment costs is probably not controversial, their exact form and precise magnitude is far from having been settled.

One's estimate of markup variations also depends on aspects of labor markets about which we are less certain. In particular, we do not know precisely how compensation of existing employees varies when there is less work to do in recessions. Indeed, one of our better estimates of this derivative of compensation with respect to productive effort (due to Bils, 1987) is based upon an estimate of the cost of adjusting employees (together with the assumption that firms minimize costs so that the cost of an additional effective hour of effort is equalized across these two margins). But perhaps the hardest problem is that, particularly outside the United States, we are not sure to what extent firms can simply take the wage that they pay per unit of effort as a parameter outside their control and to what extent this wage is the result of bargaining between workers and employees. In this latter case, the connection between the real wage and the marginal product of labor depends also on the character of this bargaining, as emphasized by Blanchard (1997) in his discussion of markups in Europe. Thus, while we have treated the ratio of the marginal product of labor to the (marginal) wage as equal to the markup, this inference is not necessarily correct if workers and firms bargain over both the level and the terms of employment. Product market considerations of the sort we have emphasized would still play a role in such a setting, but measuring the effect of these product market distortions becomes much harder.

Footnotes

1. Many authors instead define the "markup" or "price-cost margin" as $(P - MC)/P$. The two quantities are obviously monotonic transformations of one another.

2. For example, in the case of a competitive industry, the industry supply curve is simply given by $P_i = Pc(y_i)$. An increase in industry demand results in a movement up this curve, to a higher relative price as well as higher output.

3. This view of the role of markup variations in accounting for aggregate fluctuations is also one with a long history; two early proponents were Robinson (1932) and Kalecki (1938).

4. See Rotemberg and Woodford (1995) for further discussion of several of the leading models, with greater attention to the structure of general equilibrium models incorporating these mechanisms, and to issues such as calibration and numerical solution of such models.

5. Well-known proposals include nominal wage rigidity, as in Keynes (1936), as a result of which inflation lowers the real wage and hence real marginal cost; and variations in the household labor supply schedule due to wealth effects and intertemporal substitution effects, as in Barro's (1981) analysis of the effects of government purchases. Evaluation of their importance is beyond the scope of this survey, though it is important to remember that these proposals require that real wages fall (by as much as the marginal product of labor) for output to expand.

6. The crucial role of markup variations in explaining the real effects of purely nominal disturbances is stressed in particular by Kimball (1995).

7. Note, however, that the studies of individual industries discussed in section 2.4 do attempt to measure industry markups, rather than levels of real marginal cost.

8. Here and below, we use the notation F_H to mean the partial derivative of F with respect to its second argument, the effective labor input zH, rather than with respect to H.

9. Kydland and Prescott, 1988; Solon *et al.*,, 1994. This is not true of all *industry* wages, however. See Chirinko (1980), Rotemberg and Saloner (1986) and Solon *et al. (1994)*. For a review of this issue, see Abraham and Haltiwanger, 1995.

10. The ratio of price to unit labor cost is also used as an empirical proxy for the markup in studies such as Phelps (1994).

11. The denominator is thus obtained by adding depreciation (the difference between GNP and NNP) to the conventional concept of "national income".

12. Our sample stops in 1993 because, at the time these calculations were made, the pre-1960 data were not comparable to the more recent revised NIPA data. The results from 1970 onwards were the same for the two data sets, however.

13. For further discussion of the properties of this filter, see King and Rebelo (1994).

14. We also considered labor shares detrended with a linear trend. This had only a negligible effect on our results.

15. As is true of all the results of Table 2, similar results obtain when we use detrended hours as our cyclical indicator.

16. Once we depart from the assumption of constant returns to scale, it is important to distinguish between firm production functions and the relation that exists between aggregate inputs and outputs. We now assume that each firm is the sole producer of a differentiated good, so that the overhead costs cannot be reduced by simply concentrating all production in a single firm. In a symmetric equilibrium, where the same quantity is produced of each good using the same factor inputs, then this equation also indicates the relation that exists among aggregate output and aggregate factor demands.

17. There are other ways of modeling increasing returns. In particular, one might suppose that marginal cost declines with output; an econometric specification of this kind is estimated, for example, by Chirinko and Fazzari (1994). The notion that marginal cost declines with output is problematic, however. For many firms, increasing output involves an increase in either the number of machines that are employed or an increase in the number of hours for which a given number of machines are used. Both of these seem inconsistent with declining marginal cost since more efficient machines would presumably be used first. More generally, firms whose technology has a declining marginal cost over some range would benefit by bunching production so that their plants are idle some of the time, and output,

when positive, is always at a level sufficiently large that marginal cost is not declining in output.

18. Here we assume that the overhead labor requirement is acyclical. This depends upon an assumption that entry of either firms or plants is slow, as in Rotemberg and Woodford (1995) and Ambler and Cardia (1996), and so can be neglected at business cycle frequencies. The consequences of variable entry are considered further below, in section 3.

19. Rotemberg and Woodford (1991) use a variant of this method to construct series for markup changes using aggregate U.S. data. Assuming an average markup of 1.6 and an elasticity of substitution equal to one (their baseline case), they find that markups fall by about one percent when hours increase by one percent. The constructed markup series is also strongly negatively correlated with fluctuations in aggregate hours worked. Portier (1995) uses the same method on French data, and assumes an average markup of 1.373 and an elasticity of substitution equal to one. His estimates imply that a one percent increase in GDP is associated with about a 1.5 percent reduction in markups. Thus markups would appear to be more counter-cyclical for France (a finding that is especially striking given the lower assumed returns to scale).

20. One might ask, if such costs exist, why firms do not minimize costs by hiring all of the time of those employees that they hire at all. The answer must be that firms face a wage schedule that is not simply linear in the number of hours worked by a given employee, as discussed below. Note that this hypothesis about individual wages is of no consequence for the marginal cost calculation considered in this paragraph.

21. A marginal wage that is increasing in the number of hours hired is, for example, allowed for in such studies as Abel (1978), Shapiro (1986), Bils (1987), and Basu and Kimball (1994).

22. This conclusion depends upon an assumption that only person-hours enter the production function, rather than employment or hours per employee mattering separately.

23. The fact that $V(H)$ is modeled as a fraction that rises continuously with H, rather than being zero for all $H \leq 40$ hours per week and one for all $H > 40$ hours per week requires that not all employees work the same number of hours. The nature and consequences of this

heterogeneity are not explicitly modeled.

24. This average elasticity is slightly smaller than the elasticity of 1.6 indicated by the figures given in the text relating to an increase from 40 to 41 hours per week.

25. Bils studies the variations of production-worker hours in manufacturing, and computes the marginal cost of increasing output through an increase in production-worker hours only, holding other inputs fixed, including non-production-worker hours. Thus in (2.9), $\hat{s_H}$ refers to fluctuations in the share of production-worker wages. Because he assumes a production function which is isoelastic in production-worker hours, holding fixed the other inputs, $a = 0$ in his calculations.

26. His cyclical indicator is the difference between current production-worker employment and a moving average of that series. Note that Bils does not assume, as in the simple analysis above, that employment is fixed in advance and that all short-run variation in hours occurs on the hours-per-employee margin. In fact, in his "second method" of computing the cyclical variability of the marginal wage, he explicitly considers substitution between the employment and hours-per-employee margins.

27. This assumption is more appealing in the case that H is interpreted to refer solely to production-worker hours, as in Bils's (1987) work, rather than total hours.

28. In this equation, s_H refers to wH/PY as before. In order for this to correspond to labor compensation as a share of value added, one must assume that the adjustment-cost inputs are not purchased from outside the sector of the economy to which the labor-share data apply. However, to a first-order approximation, it does not matter whether the adjustment costs are internal or external, as discussed below.

29. More generally, we shall use the notation γ_{xt} to denote the growth rate x_t/x_{t-1}, for any state variable x.

30. Even though they allow for costs of changing employment, Askildsen and Nilsen (19970 do not find any industries with countercyclical markups in their study of Norwegian manufacturing industries. However, their adjustment-cost parameter is often estimated to have the wrong sign and one would expect the markups computed on the basis of these

estimates to be procyclical.

31. Bils and Cho (1994) assume a convex cost of adjusting the employee-to-capital ratio, interpreting this as a cost of changing the organization of production, rather than a cost of hiring and firing employees. Because most variations in the employment-to-capital ratio at business-cycle frequencies are due to variations in employment, the consequences of such a specification are similar to those of the more familiar assumption of convex costs of changing the number of employees.

32. Studies that estimate separate adjustment costs for variations in employment and in the number of hours worked per employee, such as Shapiro (1986), tend to find insignificant adjustment costs for hours.

33. Bils is able to estimate this equation by assuming parametric functional forms for the functions $W'(h)$ and $\phi(\gamma_N)$, and assuming that κ_t is a constant multiple of the straight-time wage. He also notes that the term w_t should refer not simply to the average hourly wage, but to total per-employee costs divided by hours per employee; the numerator thus includes the costs of other expenses proportional to employment but independent of the number of hours worked per employee, such as payments for unemployment insurance. In fact, identification of the parameters in (2.17) is possible only because w_t is assumed not to be given by a time-invariant function $W(h_t)/h_t$, but rather by $(W(h_t) + F_t)/h_t$, where the shift term F_t representing additional per-employment costs is time-varying in a way that is not a function of h_t.

34. Of the remaining hours paid for, according to survey respondents, about two-thirds represent an increase in employee time devoted to non-production tasks, while the other third represents an increase in employee time that is not used at all. Fair (1985) offers corroborating evidence.

35. Models in which output fluctuations result from changes in firms' desired markups can also explain why labor hoarding *should* be counter-cyclical, as is discussed further in section 2.3. At least some models in which fluctuations in output result from shifts in the real marginal cost schedule have the opposite implication: periods of low labor costs should

induce increases *both* in the labor force employed in current production *and* in the labor force employed in maintenance tasks.

36. For example, models of variable effort are sometimes referred to as models of "labor hoarding", as in Burnside *et al. (1993).*

37. They provide evidence of a statistical correlation between hours per worker and other proxies for capital utilization. Their econometric results are consistent with an assumption that capital utilization is proportional to hours per employee, a result that also has a simple interpretation in terms of a common work-week for all inputs. On the other hand, as Basu and Kimball (1994) note, this correlation need not indicate that firms are forced to vary the two quantities together.

38. More generally, belief that λ should take a significant positive value, perhaps on the order of one, reduces the significance of variations in η_H as a contribution to implied markup variations, since both y and h are strongly procyclical. It is not plausible, however, to suppose that λ should be large enough to make $\hat{y} - \lambda\hat{h}$ a significantly *countercyclical* factor.

39. This assumption allows for increasing returns, but requires that they take the form of increasing returns in the value-added production function $V(K, zH)$.

40. This is shown in the fourth row of his Table 5. He regresses the percentage change in m on the percentage change in Q, for each of 21 two-digit U.S. manufacturing industries. He instruments output growth using the Ramey-Hall instruments for non-technological aggregate disturbances. He also shows that intermediate inputs rise more than does a cost-weighted average of primary inputs (labor and capital), using the same instruments; as one should expect, the regression coefficient in this case is much larger.

41. The last line of his Table 5 indicates an increase in m of only 0.12 percent for each one percent increase in the relative price of primary and intermediate inputs. His estimates of the cyclicality of materials input use indicate three times as large an elasticity for M/V as for M/Q (comparing lines 2 and 4 of that table), though the estimated elasticity of M/V is reduced when labor hoarding is controlled for. This would suggest an increase in M/V of at most 0.36 percent for each percent increase in the relative price of inputs.

42. Similar issues arise with the study of Felli and Tria (1996) who use the price divided by overall average cost as a measure of the markup. They compute this by dividing total revenue by total cost including an imputed cost of capital (which depends on a measure of the real interest rate). Leaving aside the difficulties involved in measuring the cost of capital, it is hard to imagine that adding together the shares of labor, materials and capital is appropriate for computing markups unless each share in isolation is appropriate as well. In addition, the existence of adjustment costs of capital probably make the marginal cost that results from producing an additional unit by adding capital considerably more procyclical than average capital cost. These adjustment costs may also rationalize the dynamic relation they find between their ratio of average cost to output and output itself.

43. This aspect of inventory behavior has been much discussed as an embarrassment to the "production smoothing" model of inventory demand, which implies that inventories should be drawn down in booms (*e.g.*, Blinder, 1986). That prediction is obtained by adjoining to (2.25) the assumptions that b is decreasing in I and that real marginal cost is increasing in the level of production Q.

44. A theoretical rationale for this is provided in terms of a model of the stockout-avoidance demand for inventories.

45. The price data for the particular industries considered by Bils and Kahn are ambiguous in this regard; they find that (given their measures of variations in marginal cost) markups are countercyclical in some industries but procyclical in others. This means that certain of their sectors have strongly procyclical relative prices for their products – something that cannot be true of industries in general.

46. We have considered separately each of these different ways in which firms can increase their output and their associated marginal cost. An alternative is to postulate a relatively general production (or cost) function, estimate its parameters by assuming that firms minimize costs, and thereby obtain estimates of marginal cost that relate to many inputs at once. One could then compare this "average" estimate of marginal cost to the price that is actually charged. Morrison (1992) follows a related approach.

47. In taking this view, of course, we assume that variations in technical progress are essentially exogenous, at least at business-cycle frequencies.

48. Note that the latter assumption is *necessary* for equilibrium, if we assume that markups do not vary because product markets are perfectly competitive. In the case of market power but a constant markup (as in a model of monopolistic competition with Dixit-Stiglitz preferences and perfectly flexible prices – see below), a mildly increasing marginal product of labor schedule is theoretically possible, but does not seem to us appealing as an empirical hypothesis.

49. For example, Bils (1987) assumes a relationship of this kind, where w_0 represents the time-varying straight-time wage, while the function $v(H)$ reflects the nature of the overtime premium, which is time-invariant in percentage terms.

50. See Appendix 2 in Rotemberg and Woodford (1991).

51. Which case is actually correct will depend upon the relative degrees of curvature of the various schedules that enter into the right-hand sides of (2.27) and (2.28).

52. These first two series have been widely used as instruments for non-technological sources of variation in U.S. economic activity, following the precedent of Hall (1988, 1990).

53. For a recent survey, see Leeper, Sims, and Zha (1996).

54. This result is discussed extensively by Bruno and Sachs (1985), who use it to assert that the unemployment following the oil shocks was due to real wage demands being too high.

55. Finn (1997), however, finds larger declines in the case of a competitive model that allows for variable utilization of the capital stock.

56. The same explanation is offered by Clark for the behavior of the relative prices of goods at different stages of processing.

57. Interruptions of the supply of bank credit certainly can significantly affect the level of economic activity, but the most obvious channel through which this occurs is through the effects of financing costs upon aggregate demand. Financing costs are obviously important determinants of investment demand, the demand for consumer durables, and inventory ac-

cumulation; but a contraction of these components of aggregate demand can easily cause a reduction of equilibrium output, without the hypothesis of an increase in supply costs.

58. For a more elaborate analysis of the evolution of cyclical markups in four relatively narrowly defined (four digit) industries, see Binder (1995). He finds that these four industries do not have a common pattern of markup movements, though none of them has strongly countercyclical markups.

59. Because, as Hall notes, pure profits are near zero for U.S. industries, $s_K + s_H$ has a value near one for a typical industry; hence the two types of factor shares, and the two types of productivity residuals, are quantitatively similar in most cases.

60. If we imagine a competitive auction market for labor, then (3.6) is just the inverse of the labor supply curve. But a schedule of the form (3.6) is also implied by a variety of non-Walrasian models of the labor market, including efficiency wage models, union bargaining models, and so on. See, *e.g.*, Layard *et al. (1991)*, Lindbeck (1993), and Phelps (1995) for examples of discussions of equilibrium employment determination using such a schedule.

61. As noted earlier, this implies that (2.11) holds with ω replaced by $\omega(e)$.

62. Note that this follows from the fact that both equations (2.12) and (2.22) apply in this case.

63. They present their analysis as a criticism of sticky-price models of the effects of monetary policy; but in fact their criticism relates simply to the fact that the model is one in which output increases due to a reduction in markups.

64. Solon *et al.* find a considerably larger elasticity for the wage of individuals, once one controls for cyclical changes in the composition of the workforce. However, for purposes of the cyclical profits calculation, it is the elasticity of the average wage that matters; the fact that more hours are low-wage hours in booms helps to make profits more procyclical.

65. This is what Bils' estimates imply for the ratio of marginal wage to average wage when the margin in question is an increase in weekly hours per employee, and the derivative is evaluated at a baseline of 40 hours per week. (As noted above, Bils finds that this ratio rises as hours per employee increase.) In applying this ratio to equation (3.15), we assume that

the marginal cost of additional hours is the same whether they are obtained by increasing hours per employee or by increasing the number of employees, as must be true if firms are cost-minimizing.

66. Note that we do *not* here assume a structural relation between the two variables.

67. Because we later want to compute the sample values of $\hat{y}^{\mu}$ we truncate k so that it runs only between zero and eighteen. Given that our $\tilde{\lambda}$ equals .79, this truncation should not have a large effects on our results.

68. Note that we have made η_W, the elasticity of the wage with respect to hours along the aggregate labor supply curve, somewhat smaller than before because our use of a positive ψ implies that wages rise with output not only because hours rise but also because effort rises.

69. For another setting where inferences regarding markups are significantly affected by supposing that there are costs of adjusting labor, see Blanchard (1997).

70. Note that we have discussed above reasons why this need not be so, for example when a firm's marginal wage differs from its average wage. As Kimball (1995) shows, deviations from this assumption may matter a great deal for the speed of aggregate price adjustment, but we confine our presentation to the simplest case here.

71. This is what Gali (1998), Basu, Fernald and Kimball (1997), Kiley (1996) find to be true in U.S. data, using a variety of quite different methods. Shea (1998), who identifies productivity shocks from data on R&D spending and patents, does not find this contractionary effect upon input demand, though his identified shocks also have little impact on long run output.

72. See Kimball (1995) and Goodfriend and King (1997) for more detailed sketches of this program, which the latter authors term "the New Neoclassical Synthesis". Goodfriend (1997) also stresses the importance of markup variations in accounting for the real effects of monetary policy.

73. This value is about .75 for her quarterly model, which implies an average time between price changes of approximately 14 months. This represents less frequent price adjustment than is observed in most sectors of the U.S. economy, according to the survey evidence

presented in Blinder *et al. (1997)*. The coefficient κ estimated by Sbordone can be reconciled, however, with more frequent price adjustments if one hypothesizes variations in desired markups, as discussed in section 4.3.

74. This means that the model of markup variation (4.5), combined with the simple measure (2.4) of marginal costs, can account for 88% of the observed variability of the log ratio of price to unit labor cost (or equivalently, of the log labor share) over this period.

75. See Bils (1989) for a related idea.

76. See Heijdra (1995) for an analysis where government purchases may affect markups through this channel.

77. For a survey of much of this theoretical literature and its applications, see Klemperer (1995).

78. Felli and Tria (1996) argue that their proposed markup series is consistent with this implication.

79. Phelps (1994) emphasizes that this can be overturned in open economies under flexible exchange rates. Expansionary fiscal policies then tend to appreciate the exchange rate thereby forcing domestic firms to lower their markups in order to compete effectively with foreign firms.

80. The model as expounded here and in the literature, however, involves flexible prices. The extension of the theory to allow for delays in price adjustment would seem a high priority for future research.

81. Portier (1995) also considers a model where markups fall not only because entry occurs in booms but also because the threat of entry leads incumbent firms to lower their prices.

82. Thus it may help to reconcile the estimate of κ by Sbordone (1997), based on the comovement of aggregate indices of prices and labor costs, with microeconomic evidence on the frequency of price changes.

83. In focusing on the effect of markup variations (rather than the effect of the average level of the markup) we are assigning to imperfect competition a role in macroeconomics that is quite different from the one which Carlton (1996) argues is unimportant. For a discussion of

the effect of the markup level, see also Rotemberg and Woodford (1995) and the references cited therein.

6 References

Abel, Andrew B. (1978), Investment and the Value of Capital, Ph.D. dissertation, M.I.T.

Abraham, Katherine G. and John C. Haltiwanger (1995), Real Wages and the Business Cycle, Journal of Economic Literature 33:1215 1264.

Ambler, Stephen and Emanuela Cardia (1996), The Cyclical Behavior of Wages and Profits under Imperfect Competition, mimeo.

Askildsen, Jan E. and Øivind A. Nilsen (1997), Markups, Business Cycles and Factor Markets: An Empirical Analysis, University of Bergen mimeo.

Bagwell, Kyle and Robert W. Staiger (1997), Collusion Over the Business cycle, Rand Journal of Economics 28:82 106.

Ball, Laurence and David Romer (1990), Real Rigidities and the Nonneutrality of Money, Review of Economic Studies 57:183 203.

Baker, Jonathan B. and Peter A. Woodward, Market Power and the Cross-Industry Behavior of Prices around a Business Cycle Trough, mimeo 1994.

Barro, Robert J. (1981), Output Effects of Government Purchases, Journal of Political Economy 89:1086 1121.

Barro, Robert J., and Robert G. King (1984), Time Separable Preferences and Intertemporal Substitution Models of Business Cycles, Quarterly Journal of Economics 99:817 839.

Basu, Susanto (1995), Intermediate Inputs and Business Cycles: Implications for Productivity and Welfare, American Economic Review 85:512 531.

Basu, Susanto, and Miles Kimball (1994), Cyclical Productivity with Unobserved Input Variation, unpublished, Univ. of Michigan.

Benabou, Roland (1992), Inflation and Markups: Theories and Evidence from the Retail Trade Sector, European Economic Review, 36:566 574.

Basu, Susanto, John Fernald and Miles Kimball (1997), Are Technology Improvements Contractionary?, mimeo.

Bils, Mark (1987), The Cyclical Behavior of Marginal Cost and Price, American Economic Review 77:838 857.

— (1989), Pricing in a Customer Market, Quarterly Journal of Economics 104:699 718.

Bils, Mark and Jang-Ok Cho (1994), Cyclical Factor Utilization, Journal of Monetary Economics 33:319 354.

Bils, Mark and James A. Kahn (1996), What Inventory Behavior Tells us about Business Cycles, Rochester Center for Economic Research Working Paper 428.

Binder, Michael (1995), Cyclical Fluctuations in Oligopolistic Industries under Heterogeneous Information: An Empirical Analysis, mimeo.

Blanchard, Olivier (1997), The Medium Term, mimeo.

Blanchard, Olivier J. and Danny Quah (1989), The Dynamic Effects of Aggregate Supply and Demand Disturbances, American Economic Review 79:655 673.

Blinder, Alan S., Elie R.D. Canetti, David Lebow, and Jeremy B. Rudd (1998), Asking About Prices: A New Approach to Understanding Price Stickiness (Russell Sage Foundation, New York).

Boldrin, Michele and Michael Horvath (1996), Labor contracts and Business Cycles, Journal of Political Economy 103:972 1004.

Borenstein, Severin and Andrea Shepard (1996), Dynamic Pricing in Retail Gasoline Markets, Rand Journal of Economics 27:429 451.

Bruno, Michael, and Jeffrey Sachs (1985), Economics of Worldwide Stagflation, (Harvard University Press, Cambridge).

Burnside, C., Martin Eichenbaum and Sergio Rebelo (1993), Labor Hoarding and the Business Cycle, Journal of Political Economy , 101:245 273.

Calvo, Guillermo (1983), Staggered Prices in a Utility-Maximizing Framework, Journal of Monetary Economics 12:383 398.

Campbell, John Y. (1987), Does Saving Anticipate Declining Labor Income? An Alternative Test of the Permanent Income Hypothesis, Econometrica 55:1249 1273.

Carlton, Dennis (1996), A Critical Assessment of the Role of Imperfect Competition in Macroeconomics, NBER Working Paper 5782.

Chatterjee, Satyajit, Russell Cooper and B. Ravikumar (1993), Strategic Complementarity in Business Formation: Aggregate Fluctuations and Sunspot Equilibria, Review of Economic Studies 60:795 812.

Chevalier, Judith A. and David Scharfstein (1995), Liquidity Constraints and the Cyclical Behavior of Markups, American Economic Review Papers and Proceedings 85:390 396.

— and — (1996), Capital-Market Imperfections and Countercyclical Markups: Theory and Evidence, American Economic Review , 86:703 725.

Chirinko, Robert S. (1980), The Real Wage Rate over the Business Cycle, Review of Economics and Statistics 62:459 461.

Chirinko, Robert S. and Stephen M. Fazzari (1994), Economic Fluctuations, Market power and Returns to Scale: Evidence from Firm-Level Data, Journal of Applied Econometrics 9:47 69.

— and — (1997), Market Power, Inflation, and Product Market Structure, mimeo, Emory University.

Christiano, Lawrence J., Martin Eichenbaum and Charles L. Evans (1996), Sticky Price and Limited Participation Models of Money: A comparison, NBER Working paper 5804.

Cochrane, John and Argia Sbordone (1988), Multivariate Estimates of the Permanent Components of GNP and Stock Prices, Journal of Economic Dynamics and Control 12:255 296.

Clark, Todd E. (1996), The Response of Prices at Different Stages of Production to Monetary Policy Shocks, mimeo.

Devereux, Michael B., Allen Head and Beverly J. Lapham (1996), Monopolistic Competition, Increasing Returns and the Effects of Government Spending, Journal of Money, credit and Banking 28:233 254.

Dixit, Avinash and Joseph Stiglitz (1977), Monopolistic Competition and Optimum Product Diversity, American Economic Review 67:297 308.

Domowitz, Ian, R. Glenn Hubbard and Bruce C. Petersen (1986), Business Cycles and the Relationship Between Concentration and Price-Cost Margins, Rand Journal of Economics 17:1 17.

—, — and — (1987), Oligopoly Supergames: Some Empirical Evidence on Prices and Margins, Journal of Industrial Economics 35:379 398.

—, — and — (1988), Market Structure and Cyclical Fluctuations in U.S. Manufacturing, Review of Economics and Statistics 70:55 66.

Dunlop, John T. (1938), The Movement in Real and Money Wage Rates, Economic Journal 18:413 434.

Eichenbaum, Martin (1989), Some Empirical Evidence on the Production Level and the Production Cost Smoothing Models of Inventory Investment, American Economic Review 79:853 864.

Ellison, Glenn (1994), Theories of Cartel Stability and the Joint Executive Committee, Rand Journal of Economics 25:37 57.

Evans, George W. (1989), Output and Unemployment Dynamics in the United States, 1950-1985, Journal of Applied Econometrics 4:213 237.

Evans, George W. and Lucrezia Reichlin (1994), Information, Forecasts and Measurement of the Business Cycle, Journal of Monetary Economics 33:233 254.

Farrell, Joseph and Carl Shapiro (1988), Dynamic Competition with Switching Costs, Rand Journal of Economics 19:123 137.

Fay, Jon A. and James L. Medoff (1985), Labor and output over the Business cycle, American Economic Review 75:638 655.

Felli, Ernesto and Giovanni Tria (1996), Markup Pricing Strategies and the Business Cycle, University of Rome, mimeo.

Finn, Mary (1997), Perfect Competition and the Effects of Energy Price Increases on Economic Activity, mimeo.

Galeotti, Marzio and Fabio Schiantarelli (1998), The Cyclicity of Markups in a Model with Adjustment Costs: Econometric Evidence for U.S. Industry, Oxford Bulletin of Economics and Statistics, forthcoming.

Gali, Jordi (1994), Monopolistic Competition, Business Cycles, and the Composition of Aggregate Demand, Journal of Economic Theory, 63:73 96.

— (1998), Technology, Employment and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?, mimeo.

Gomme, Paul and Jeremy Greenwood (1995), On the Cyclical Allocation of Risk, Journal of Economic Dynamics and Control 19:91 124.

Goodfriend, Marvin (1997), A Framework for the Analysis of Moderate Inflations, Journal of Monetary Economics 39:45 65.

Goodfriend, Marvin, and Robert G. King, The New Neoclassical Synthesis and the Role of Monetary Policy, NBER Macroeconomics Annual, 231 282.

Gottfries, Nils (1986), Price Dynamics of Exporting and Import-Competing Firms, Scandinavian Journal of Economics 88:417 436.

— (1991), Customer Markets, Credit Market Imperfections and Real Price Rigidity, Economica 58:317 323.

Greenwald, Bruce, Joseph Stiglitz and Andrew Weiss (1984), Informational Imperfections in the Capital market and Macroeconomic Fluctuations, American Economic Review Papers and Proceedings 74:194 199.

Hall, Robert E. (1980), Employment Fluctuations and Wage Rigidity, Brookings Papers on Economic Activity, 91 123.

Hall, Robert E. (1988), The Relation Between Price and Marginal Cost in U.S. Industry, Journal of Political Economy 96:921 948.

— (1990), Invariance Properties of Solow's Productivity Residual, in Peter A. Diamond ed., Growth, Productivity and Unemployment, Essays to Celebrate Bob Solow's Birthday (MIT Press, Cambridge).

Hamilton, James D. (1983), Oil and the Macroeconomy Since World War II, Journal of Political Economy 91:228 248.

Hansen, Gary D. and Thomas J. Sargent (1988), Straight Time and Overtime in Equilibrium, Journal of Monetary Economics 21:281 308.

Heijdra, Ben J. (1995), Fiscal Policy Multipliers: The Role of Market Imperfection and Scale Economies, mimeo, University of Amsterdam.

Hultgren, Thor(1965), Costs, Prices and Profits: Their Cyclical Relations, (National Bureau of Economic Research, New York).

Kalecki, Michael (1938), The Determinants of the Distribution of National Income, Econometrica 6:97 112.

Keynes, John Maynard (1936), The General Theory of Employment, Interest and Money (Macmillan, London).

— (1939), Relative Movements of Real Wages and Output, Economic Journal 49:34-51.

Kiley, Michael T. (1996), Labor productivity in U.S. Manufacturing: Does Sectoral Co-

movement Reflect Technology Shocks?, Federal Reserve Board, mimeo.

— (1997), Staggered Price Setting and Real Rigidities, Federal Reserve Board, mimeo.

Kimball, Miles S. (1995), The Quantitative Analytics of the Basic Neomonetarist Model, Journal of Money Credit and Banking 27:1241 1277.

King, Robert and Sergio Rebelo (1993), Low Frequency Filtering and Real Business Cycles, Journal of Economic Dynamics and Control 17:207 231.

Klemperer, Paul D. (1987), Markets with Consumer Switching Costs, Quarterly Journal of Economics 102:375 394.

— (1995), Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics and International Trade, Review of Economic Studies 62:515 539.

Kollman, Robert (1996), The Cyclical Behavior of Markups in U.S. Manufacturing and Trade: New Empirical Evidence Based on a Model of Optimal Storage, mimeo, Univ. of Montreal.

Kydland, Finn E. and Edward C. Prescott (1982), Time to Build and Aggregate Fluctuations, Econometrica 50:1345 1370.

— and — (1988), Cyclical Movements of the Labor Input and its Real Wage, Working Paper 413, Research Department, Federal Reserve Bank of Minneapolis.

Layard, Richard, Stephen Nickell, and Richard Jackman (1991), Unemployment (Oxford Univ. Press, Oxford).

Leeper, Eric M., Christopher A. Sims, and Tao Zha (1996), What Does Monetary Policy Do?, Brookings Papers on Economic Activity, 1 63.

Lindbeck, Assar (1993), Unemployment and Macroeconomics (MIT Press, Cambridge).

Lucas, Robert E. Jr., Capacity (1970), Overtime and Empirical Production Functions, American Economic Review Papers and Proceedings 60:23 27.

Mankiw, N. Gregory, Julio J. Rotemberg and Lawrence Summers (1985), Intertemporal Substitution in Macroeconomics, Quarterly Journal of Economics 100:225 251.

Mills, Frederick (1936), Prices in Recession and Recovery, (National Bureau of Economic Research, New York).

Mitchell, Wesley Clair (1941), Business Cycles and their Causes , (University of California Press, Berkeley).

Morrison, Catherine J. (1992), Markups in U.S. and Japanese Manufacturing: A Short-Run Econometric Analysis, Journal of Business and Economic Statistics, 10:51 63.

Murphy, Kevin M., Andrei Shleifer and Robert W. Vishny (1989), Building Blocks of Market Clearing Business Cycle Models, NBER Macroeconomics Annual, 247-86.

Parker, Jonathan A. (1996), The Timing of Purchases, Market Power and Economic Fluctuations, mimeo.

Phelps, Edmund (1994), Structural Slumps (Harvard University Press, Cambridge).

Phelps, Edmund S. and Sidney G. Winter (1970), Optimal Price Policy under Atomistic Competition, in E. Phelps ed., Microeconomic Foundations of Employment and Inflation Theory (W. W. Norton and Co., New York).

Plosser, Charles, I. (1989), Understanding Real Business Cycles, Journal of Economic Perspectives, 3:51 78.

Portier, Franck (1995), Business formation and Cyclical Markups in the French Business Cycle, Annales d'Économie et de Statistique , 37:411 440.

Prescott, Edward (1973), Efficiency of the Natural Rate, Journal of Political Economy 83:1229 1236.

Ramey, Valerie A. (1991), Non-Convex Costs and the Behavior of Inventories, Journal of Political Economy , 99:306 334

— and Matthew D. Shapiro (1998), Costly Capital Reallocation and the Effects of Government Spending, Carnegie-Rochester Conference on Public Affairs , forthcoming.

Robinson, Joan (1932), The Economics of Imperfect Competition, (Macmillan, London).

Rotemberg, Julio J. (1982), Sticky Prices in the United States, Journal of Political Economy 90:1187 1211.

Rotemberg, Julio J. and Garth Saloner (1986), A Supergame-Theoretic Model of Price Wars during Booms, American Economic Review 76:390 407.

Rotemberg, Julio J. and Michael Woodford (1991), Markups and the Business Cycle, NBER Macroeconomics Annual 63 129.

— and — (1992), Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity, Journal of Political Economy 100:1153 1207.

— and — (1995), Dynamic General equilibrium Models with Imperfectly Competitive Product Markets, in Thomas F. Cooley, ed. Frontiers of Business Cycle Research (Princeton University Press, Princeton).

— and — (1996a), Real Business Cycle Models and the Forecastable Movements in Output, Hours and Consumption, American Economics Review 86:71 89.

— and — (1996b), Imperfect Competition and the Effects of Energy Price Increases on Economic Activity, Journal of Money, Credit and Banking 28:549 577.

Sbordone, Argia M. (1996), Cyclical Productivity in a Model of Labor Hoarding, Journal of Monetary Economics 38:331 362.

— (1997), Prices and Unit Labor Costs: Testing Models of Pricing Behavior, mimeo, Princeton University.

Shapiro, Matthew D. (1986), The Dynamic Demand for Capital and Labor, Quarterly Journal of Economics 101:513 542.

Shea, John (1998), What Do Technology Shocks Do?, NBER Macroeconomics Annual, forthcoming.

Solon, Gary, Robert Barsky and Jonathan A. Parker (1994), Measuring the Cyclicality of Real Wages: How Important is Composition Bias, Quarterly Journal of Economics 109:1 26.

Tarshis, Lorie (1939), Changes in Real and Money Wage Rates, Economic Journal 19:150 154.

Taylor, John B., Temporary Price and Wage Rigidities in Macroeconomics: A Twenty-Five Year Review, this Handbook.

Woglom, Geoffrey (1982), Underemployment Equilibrium with Rational Expectation, Quarterly Journal of Economics 97:89 107.

Table 1
Correlations of Selected Variables with Cyclical Indicators

	Predicted Declines in GDP	H-P Filtered GDP	Linearly Detrended Hours	H-P Filtered Hours
Share of Compensation in After Indirect Tax Gross Product				
Long Sample				
Overall	-0.070	-0.095	0.055	-0.023
Corporate	-0.080	-0.188	0.031	-0.044
Nonf. Corp.	-0.014	-0.158	0.072	-0.014
Private	-0.009	-0.192	0.178	-0.192
Sample: 69:1-93:1				
Overall	-0.230	-0.403	-0.189	-0.015
Corporate	-0.077	-0.273	-0.010	0.064
Nonf. Corp.	0.066	-0.169	0.103	0.184
Private	-0.334	-0.466	-0.293	-0.156
Correlation of Private Labor Share with leads and lags of Cyclical Indicator				
Lead six quarters	-0.437	-0.108	-0.218	-0.136
Lead five quarters	-0.521	-0.176	-0.312	-0.211
Lead four quarters	-0.579	-0.270	-0.406	-0.276
Lead three quarters	-0.5828	-0.360	-0.461	-0.314
Lead two quarters	-0.564	-0.429	-0.454	-0.304
Lead one quarters	-0.509	-0.477	-0.407	-0.256
Lagged one quarter	-0.157	-0.283	-0.100	0.015
Lagged two quarters	-0.026	-0.110	0.063	0.162
Lagged three quarters	0.075	0.023	0.180	0.270
Lagged four quarters	0.149	0.138	0.258	0.346
Lagged five quarters	0.177	0.194	0.303	0.388
Lagged six quarters	0.213	0.222	0.317	0.406

Note: The long sample for all correlations except those involving either the labor share in the private sector or predicted declines in private GDP is 1947:1 to 1993:1. The sample for the correlations involving the labor share in the private sector starts in 1952:1. That for the correlations of predicted output declines with the other labor share starts in 1948:3 because these predicted declines are drawn from Rotemberg and Woodford (1996). The correlations with leads and lags of output are based on data from 69:1 to 93:1

Table 2
Correlation of Markup Based on Private Labor Share with Leads and Lags of Expected Declines in GDP

	a=-.4, b, c=0	b=-.4, a, c=0	c=4,a, b=0	c=8, a, b=0
Lead six quarters	0.355	0.370	0.136	-0.058
Lead five quarters	0.323	0.342	0.067	-0.169
Lead four quarters	0.256	0.289	-0.048	-0.316
Lead three quarters	0.135	0.189	-0.203	-0.478
Lead two quarters	-0.001	0.075	-0.321	-0.594
Lead one quarters	-0.163	-0.050	-0.418	-0.670
Contemporaneous	-0.402	-0.212	-0.372	-0.542
Lagged one quarter	-0.504	-0.312	-0.235	-0.319
Lagged two quarters	-0.522	-0.344	-0.162	-0.181
Lagged three quarters	-0.503	-0.337	-0.124	-0.095
Lagged four quarters	-0.451	-0.301	-0.066	-0.001
Lagged five quarters	-0.355	-0.226	-0.003	0.079
Lagged six quarters	-0.278	-0.164	0.011	0.110

Markup is based on (2.14) and (2.15) and uses the labor share in the nonfinancial corporate business sector.

Table 3 Fractions of the Variance of y Accounted by $\hat{y}^{\mu}$ and y^{*}

	$\frac{\text{Var}\Delta y^{*}}{\text{Var}\Delta y}$	$\frac{\text{Var}\Delta \hat{y}^{\mu}}{\text{Var}\Delta y}$	$\frac{\text{Var}\Delta \hat{y}^{\mu} \text{ orthogonal to } \Delta y^{*}}{\text{Var}\Delta y}$
b=-.4, a, c=0			
Innovation variances	1.43	0.05	0.01
Revisions over 5 quarters	0.88	0.06	0.06
Revisions over 9 quarters	0.86	0.08	0.08
Revisions over 13 quarters	0.90	0.07	0.07
Revisions over 17 quarters	0.90	0.05	0.05
Revisions over 21 quarters	0.91	0.05	0.05
Revisions over 25 quarters	0.91	0.04	0.04
c=8 , a, b=0			
Innovation variances	2.38	2.89	0.97
Revisions over 5 quarters	0.55	1.28	0.97
Revisions over 9 quarters	0.65	1.13	0.89
Revisions over 13 quarters	0.66	1.13	0.90
Revisions over 17 quarters	0.61	1.03	0.86
Revisions over 21 quarters	0.59	0.91	0.81
Revisions over 25 quarters	0.58	0.81	0.75
Revisions over 81 quarters	0.84	0.21	0.21

Calculations based on projecting $(w - y)$ on Z for period 1969:1-1993:1 and using properties of stochastic process in (3.21) where this stochastic process is estimated from 1948:3 to 1993:1.

Figure 1
The Evolution of Various Labor Shares

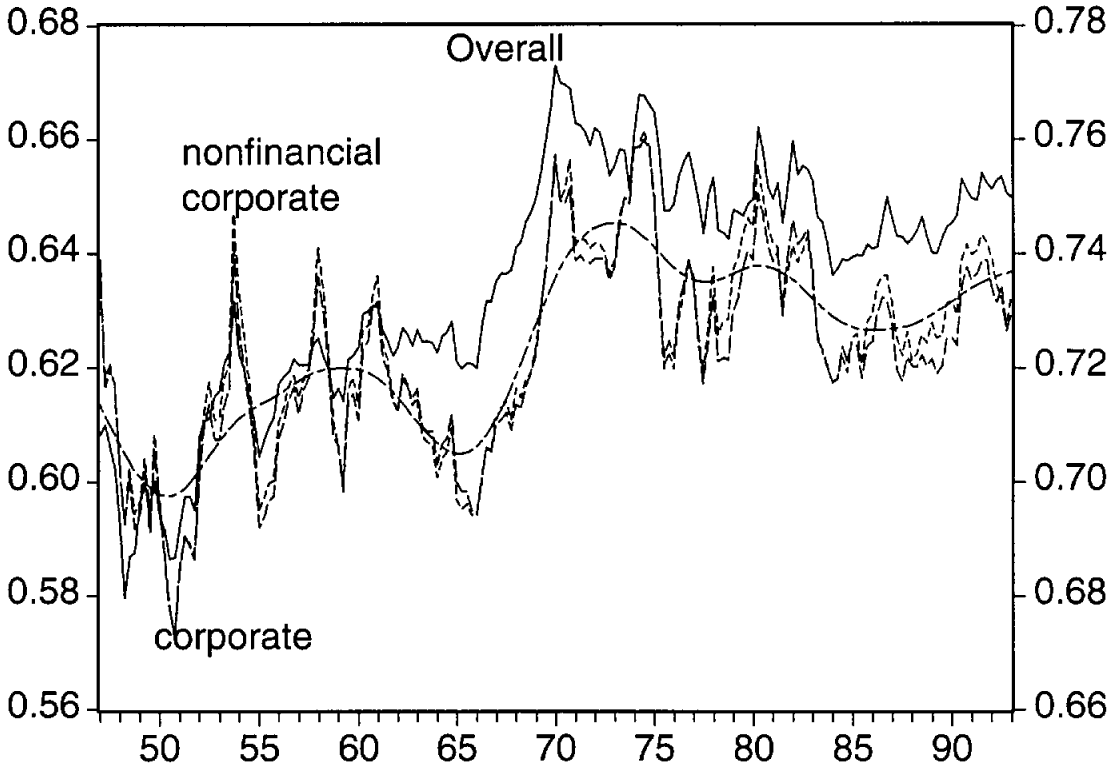

Figure 2
Labor Share in Nonfinancial Corporate
Sector and NBER Recessions

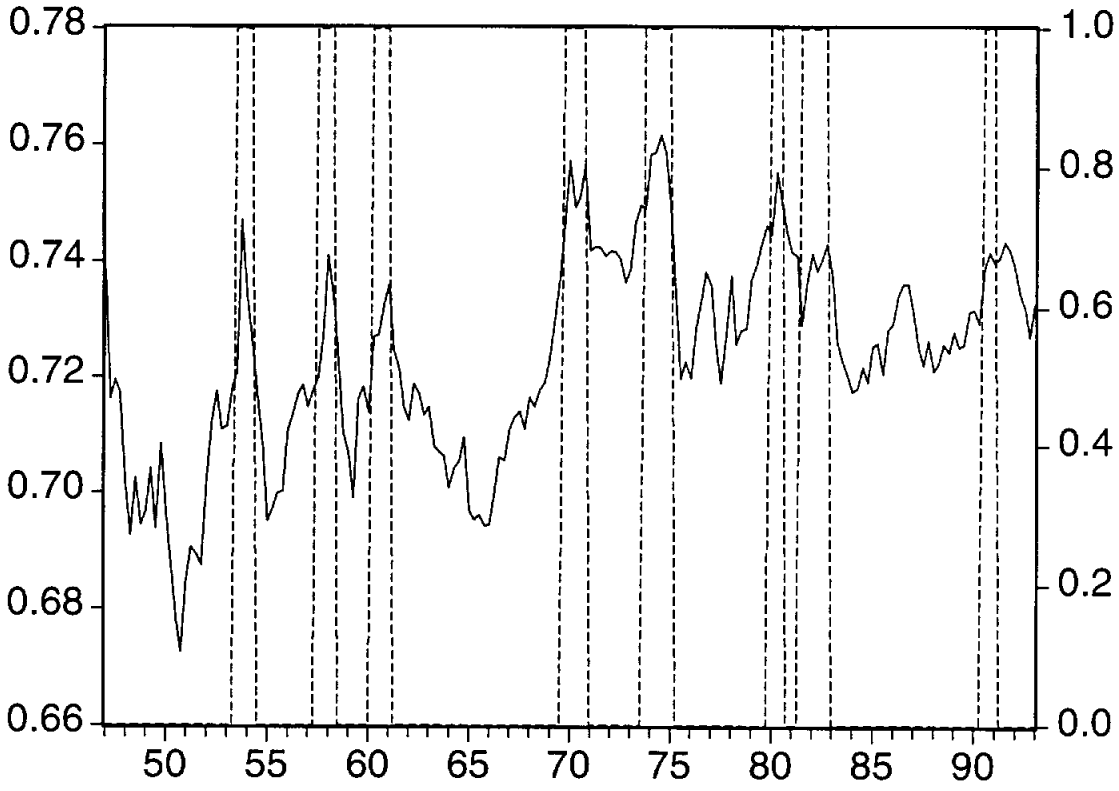

Figure 3
Constant-Markup and Actual Output

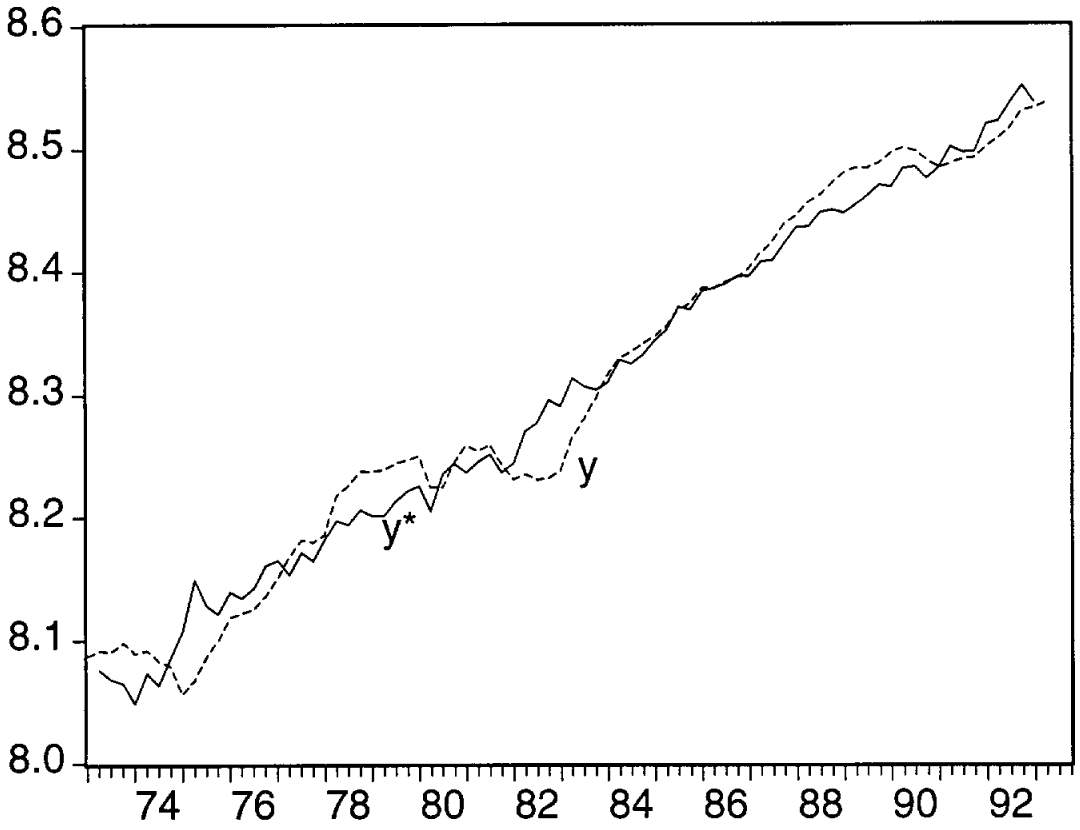

Figure 4
Markup-Induced Output Gap and
Predicted Output Declines