Expositional Essays in Macroeconomics

John E. Floyd
University of Toronto

November 30, 2009

PREFACE

This collection of three essays brings together some material I have written for my intermediate and advanced students in macroeconomics. I would like to thank my colleague Allan Hynes for helpful comments on many of the issues discussed here.

These essays are made available in the hope that others will find them useful. I will correct any errors and give credit to anyone who finds them. Should other instructors wish to modify any of the exposition here for use in their classes, or add essays on additional topics, I would be happy to make my LaTeX code available to them under appropriate conditions with respect to ensuring that those revisions and extensions also be freely available for further modification by others.

Contents

ESSAY 1

Auction, Search and Contract Theories of Unemployment

1.1 Introduction

While the traditional classical model assumed that wages adjust instantaneously to excess demand and supply, everyone knew that unemployment is a characteristic of depressed periods. Various loosely formulated arguments, most of them familiar today, emerged over the years to explain this phenomenon. Keynes, spurred by the need to explain the massive unemployment of the Great Depression, adopted the extreme assumption that money wages are rigidly fixed when aggregate demand falls below the full employment level. This view that for institutional and other reasons money wages do not respond to excess supply in the labour market dominated professional thinking for two or three decades. Yet everyone knew that this was an extreme assumption—money wages frequently fall during recessions.

In 1958, A.W. Phillips drew attention to a phenomena that had been observed from time to time for generations, namely, that when the rate of increase in money wages is high the unemployment rate tends to be low, and vice versa.[1] This observed negative relation between inflation (of prices as well as wages) and unemployment became known as the Phillips curve. It led many economists to believe that policy makers face a tradeoff between inflation and unemployment—that by tolerating a higher inflation rate, they

[1]A. W. Phillips, "The Relationship Between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861–1957," *Economica*, Nov. 1958.

can achieve a reduction in the unemployment rate. Since most people agreed that unemployment was more odious than inflation, the implication was that an inflationary monetary policy would be socially beneficial. Subsequent empirical research found the case for observed Phillips curve relationships extending over long periods to be extremely shaky.

The problem with interpreting evidence on the Phillips curve is that neither the classical nor the Keynesian models incorporate a useful theory of how wages and prices change. Wages and prices change in the Keynesian theory, if at all, as a result of institutional circumstances beyond the scope of conventional economic theory. The classical model, on the other hand, incorporates a theory of price adjustment that has wages and prices responding instantaneously to excess demand and supply in perfectly competitive markets. Not only does this latter theory fail to explain observed unemployment, it implies that labour and commodity markets are auction markets. In the absence of an auctioneer calling out prices and taking bids, there is no mechanism by which prices can change—the traditional perfect competition assumption that all buyers and sellers are too insignificant to influence price and therefore take it as given eliminates any role for individual suppliers and demanders in the price adjustment process. Of course, this fiction that prices change as a result of some implicit auction process is useful for many standard price theory problems that concern the economic forces determining relative prices. It is not useful in the present situation, however, where the key issue is the process by which prices change.

Obviously, if individual buyers and sellers are to make decisions about when and how much to change prices, most markets must be assumed to be imperfectly competitive to some degree. Since most products are differentiated in some way from their competitors and the labour services provided by each worker are to some extent unique, this assumption is easy to swallow. Only the very few markets with auctioneers or substitute auction type processes are perfectly competitive.

What then is the process by which wages and prices change in imperfectly competitive markets? Is it possible in such markets for rational workers and firms to set wages which will result in less labour being employed than would, in retrospect, be optimal? The conclusion that emerges from the analysis of these questions over the past thirty years is that there is no one single process by which prices are set. Three different wage and price setting mechanisms which can result in observed unemployment are examined in this essay—an auction model and versions of search and contract theories. The purpose is to leave students with an understanding of the processes by which wages are established and the roles of imperfect information, nonhomogeneity of skills and jobs and mutually beneficial agreements between firms and workers in wage setting. A couple of additional perspectives on the relationship between wage-setting and less-than-full employment—efficiency wage and insider-outsider theories—are not discussed here because a useful background can be obtained from assigned readings.[2]

1.2 The Auction Model

The first theory we examine is an extension of the perfectly competitive auction model. The ideas here trace back to a classic article by Robert Lucas.[3] The analysis supposes that markets are sufficiently competitive for prices to adjust to supply and demand changes almost immediately. Any worker wanting employment can find it at the prevailing wage rate or slightly less. It is then argued that because of poor information, misperceptions about the true equilibrium levels of wages and prices in relation to current levels cause workers to work harder in boom periods when wages and prices are relatively high and use the extra funds to finance leisure in slack periods when wages and prices are low. In the simplest version of this theory, the level of the upward sloping supply curve of labour (with the money wage rate on the vertical axis) is assumed to depend on expected prices while the level of the negatively sloped demand for labour by firms is assumed to depend on current prices. An unexpected expansion of aggregate demand causes firms to demand more labour, raising money wages and current prices relative to workers' price level expectations. This happens because workers mistakenly view the expansion of their own industry as a local rather than economy-wide event. Faced with what is perceived as a temporary increase in their real wage rate (temporary because entry of workers will eventually drive the wage rate down to the level existing elsewhere in the economy), workers substitute future leisure for current leisure and work more. Similarly, an unexpected decline of aggregate demand causes actual wages and prices to fall relative to expected wages and prices, leading workers to substitute leisure for work in the current period, expecting to increase work relative to leisure in subsequent periods when the wage rate has increased back to normal levels. In general, therefore, the observed unemployment rate will

[2]See George Akerlof and Janet Yellen, eds. *Efficiency Wage Models of the Labour Market*, Cambridge University Press, 1986, Chapter 1, and Assar Lindbeck, *Unemployment and Macroeconomics*, MIT Press, 1993.

[3]See Robert E. Lucas Jr., 'Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, 63, June 1973, pp. 326–334.

be below normal when prices are unexpectedly high and above normal when prices are unexpectedly low.

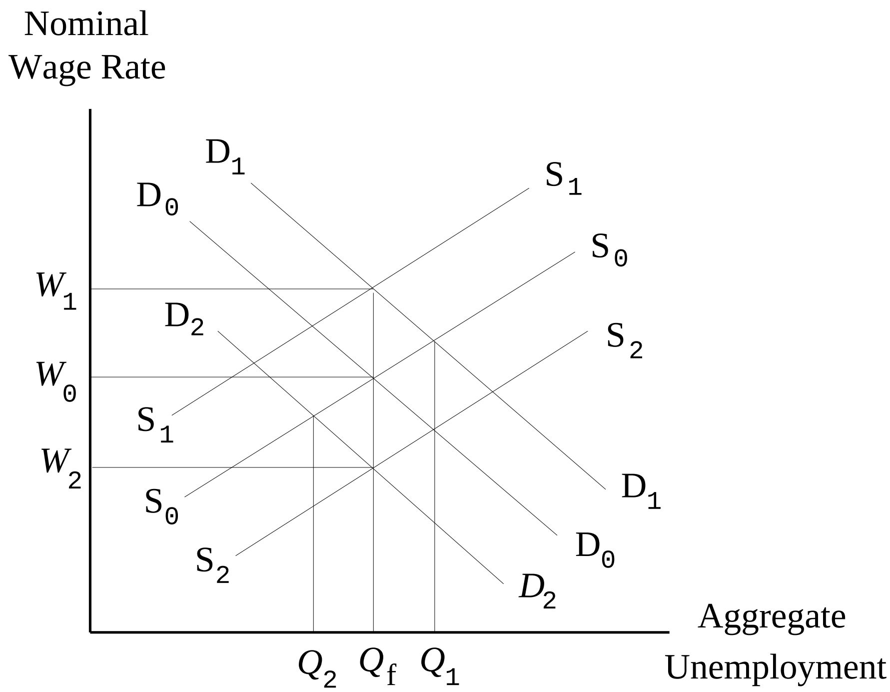

Figure 1: An Auction Model of Wages and Employment.

These ideas are illustrated in Figure 1. Aggregate employment is on the horizontal axis and the money wage rate is on the vertical one. Suppose that, starting from an initial demand curve for labour D_0D_0, there is an increase in nominal aggregate demand for output in the economy that shifts this demand curve to D_1D_1. If workers are aware that the aggregate demand shift is economy wide, they will raise their asking wages to W_1 and the supply curve of labour will shift from S_0S_0 to S_1S_1. Aggregate employment will remain at Q_f. Suppose, however, that workers do not realize that the shift in the demand curve for labour is resulting from an increase in the level of aggregate demand in the economy, but think that it represents an increase in the demand for the output of their industry alone. Nominal wage increases will be viewed as real wage increases. S_0S_0 is a short-run supply curve of labour. Workers will anticipate that the increase in the real wage rate that will result from the upward shift of the demand curve along this supply curve at constant prices elsewhere in the economy will be temporary—the wage will fall back to W_0 as new workers enter the industry and the short-run supply curve of labour shifts to the right. As the positive slope of S_0S_0 indicates, they will respond to this temporary increase in their perceived real wage rate by working an additional amount now, expecting to take an equivalent amount of time off when the wage rate falls back to W_0. The level of aggregate employment will thus rise above its normal full employment level to Q_1. The demand curve for labour expresses the quantity of labour demanded as a function of the nominal wage rate and the actual level of prices in the economy, while the supply curve for labour expresses the quantity of labour employed as a function of the nominal wage rate and the expected price level in the economy. The supply curve of labour fails to shift up to S_1S_1 in response to the economy-wide increase in aggregate demand because the workers do not perceive this increase as an economy-wide increase that will be reflected in the general price level but regard it as peculiar to their particular industries. As a result they expect the price level in the economy to remain as it was before and see no need to increase their reservation wage rate. By not recognizing that the price level in the economy as a whole and the cost of living have risen, workers have inadvertently taken a cut in their real wage rate. This causes firms to increase the quantity of labour hired, since firms will hire labour to the point where the marginal product of labour equals the real wage rate.

The situation is similar for a decline in aggregate demand that shifts the demand curve for labour to D_2D_2. Were workers to correctly perceive the situation, the supply curve would shift down to S_2S_2, the wage rate would fall to W_2, and the level of employment would remain at Q_f. Since workers do not realize that aggregate demand has fallen, however, and think the shift to D_2D_2 is specific to their industry, they will take advantage of the temporarily lower wage rate (pending exit of workers from the industry in the long-run) by working less now in hopes of making up the lost time at the wage rate W_0 in the future. The level of employment will fall to Q_2. Because unbeknownst to workers the price level in the economy as a whole has fallen—the real wage rate has gone up.

While this theory is a useful one in explaining the work-leisure tradeoffs of females with children and other part-time workers, it is inconsistent with the fact that quit rates (the fraction of workers quitting their jobs) decline in recession periods and increase during booms. The theory predicts falling nominal wages during recession periods with workers quitting (i.e., taking leisure) as a result. It also fails to explain the fact that in recession periods firms refuse to hire workers who would be willing to work for them at the going wage rate or less.

1.3 A Search Theory

Our second theory of wage adjustment is a search theory.[4] This theory starts with the proposition that unemployed workers are seeking employment, and focuses on the length of time it takes them on average to find work. The typical unemployed worker's skills and work habits are similar to but not identical with those of other competing workers. A wide range of possible employment situations are feasible to him, but both he and the firms that might potentially hire him are not fully informed of each other's relevant characteristics. Acquisition of this information requires a search process of some sort. One such process would have firms advertising vacancies and workers shopping around for jobs that have the desired characteristics and pay their reservation wages. Less frequently, one might find workers advertising their availability and firms shopping for employees.

Asking wages exist for which a worker will find employment almost immediately, but these wages will be low in comparison to what many firms would be willing to pay for his services. On the other hand, at a very high asking wage a firm could potentially be found that would offer employment, but such firms are few and far between and the worker can expect to wait a long time on average before encountering one of them. So the typical unemployed worker faces a choice—the higher his asking or reservation wage, the longer the length of time it will take him on average to find employment.

This situation for a group of workers of a particular type, without jobs and seeking work, is portrayed in Figure 2. The line $\mathrm{A_0A_0}$ gives the relationship between the reservation wage and the mean, average, or expected time it will take a worker of this type to find a job. Because the time it takes to find employment at each reservation wage is a random or stochastic variable, the actual waiting time will almost always differ from the mean or expected waiting time. The curve *abc* gives the frequency distribution or probability density function on waiting time at the wage rate W_0. Similar density functions exist for each possible wage rate. Two of these are shown on the graph, one for the wage rate W_1 and one for the wage rate W_2. The density function *abc* is drawn such that the area under it and over the straight line *ac* equals unity. All other similar density functions can be drawn equivalently.

[4] A more rigorous presentation of this theory can be found in Donald F. Gordon and Allan Hynes, "On the Theory of Price Dynamics," in Edmund S. Phelps, ed., *Microeconomic Foundations of Inflation and Employment Theory*, W.W. Norton, 1970, 369–393.

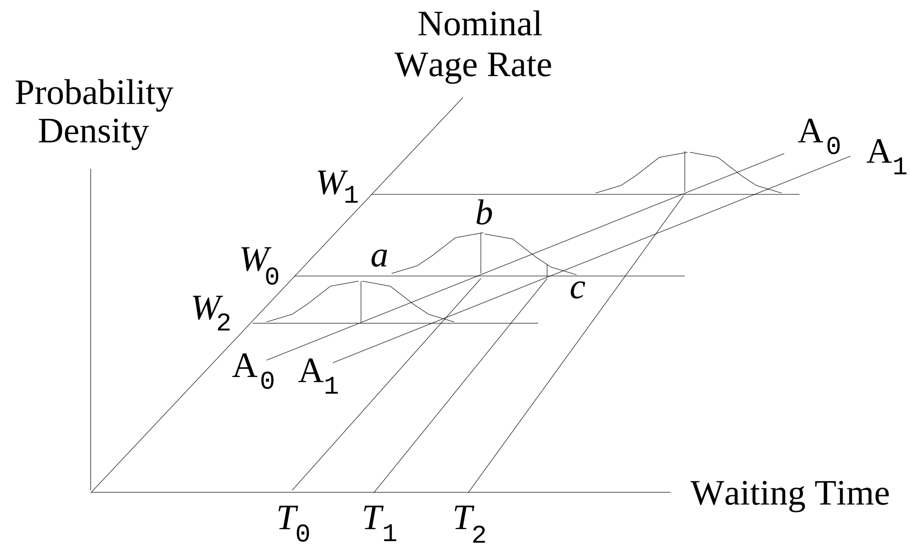

Figure 2: A Search Model of Wages and Employment.

The curve *abc* is also shown in Figure 3, which is a blow-up of the relevant area in Figure 2. The portion of the area under the curve to the left of a given waiting time gives the probability that the worker will take less than that amount of time to find a job. The probability that the waiting time at a reservation wage of W_0 will be less than T_2 is 1.0—i.e., there is a 100 percent chance that he will find a job in less time than T_2 so all the area lies to the left of T_2. The waiting time to find a job at the wage rate W_0 will be less than T_0 about half the time, since the area under the curve to the left of T_0 (the light shaded portion) is about one half. The waiting time to find a job will be less than T_1 about 95 percent of the time and greater than T_1 about 5 percent of the time—about 5% of the area under the curve (the heavily shaded portion) lies to the right of T_1 and about 95% of the area lies to the left of it. The mean or expected time to find a job will be T_0 because about half the time it will take longer than T_0 and half the time it will take less than T_0. Since the density (distance between *abc* and *ac*) greatest at waiting time T_0, this is also the most likely waiting time to occur. In short, the time it takes to find a job at the wage W_0 is likely to be near T_0, with deviations just as likely to be in the direction of a longer time than a shorter one, and with the likelihood of larger deviations in either direction from T_0 being much less than the likelihood of smaller deviations.

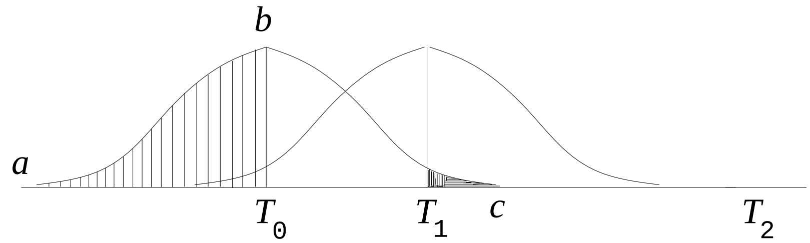

Figure 3: A Partial Blow-up of Figure 2.

Standard economic analysis tells us that the worker will react to this trade-off between wage achieved and expected waiting time by choosing the asking wage and expected waiting time combination that will maximize his expected utility. Let us suppose that the utility maximizing expected waiting time turns out to be T_0. Associated with the equilibrium expected waiting times of an aggregate of workers is what might be called the natural or equilibrium rate of unemployment, representing the normal or equilibrium proportion of workers that happen to be between jobs.

Note that T_0 is the expected waiting time—the average waiting time that will occur for a large number of workers of similar type, each specifying a reservation wage of W_0. For a single worker, or small sample of workers, the observed waiting time may be greater or less than T_0. If everyone knows conditions in the market, and the line A_0A_0 thus represents not only workers' expected waiting times at various wages but also the actual average waiting times that will be experienced for a large number of workers, then the difference between the waiting time experienced by a small group of workers and T_0 is simply one of the vagaries of life. Nothing can or should be done about it.

On the other hand, suppose that there is a possibility that market conditions may have deteriorated to the extent that the combinations of reservation wage and waiting time might now be represented by the line A_1A_1. An average waiting time of T_1 for an observed small sample of workers could represent the new mean waiting time if the tradeoff line has shifted from A_0A_0 to A_1A_1, or it could be a very unlikely though quite possible event if the tradeoff line has in fact remained at A_0A_0. In a world where information is not perfect, workers will never know for sure which is the case. If the sample is small and the deviation from the expected waiting time along the tradeoff curve on which the equilibrium is based is small, then there will be little reason to expect that the tradeoff line has shifted. When the deviation persists over a large sample of observations, however, there develops a stronger and stronger basis for concluding that the distribution has shifted. When the mean tradeoff line has shifted downward unbeknownst to the workers, unemployed workers will take longer than the normal time to find jobs, and the rate of unemployment will rise above the natural rate. Conversely, of course, if the mean tradeoff line shifts unexpectedly to the left, unemployment will fall below its natural rate.

The condition for normal unemployment rates—loosely referred to as full-employment—to occur, therefore, is that workers have good information about the state of aggregate demand (and hence about the demand for labour). When the information possessed by workers about market conditions is correct, the level of unemployment will gravitate to the natural rate. When workers overestimate aggregate demand conditions, they price themselves out of the market, causing the time taken to find jobs and the pool of workers between jobs to be unusually large. When they underestimate aggregate demand conditions, they under-price themselves. Jobs are found much more quickly than expected and the pool of unemployed shrinks below its natural size.

Once workers realize that the state of demand has shifted, of course, reservation wages will adjust and the unemployment rate will return to its natural level. Indeed, deviations of employment from the natural rate can be viewed as an integral part of the process by which workers acquire information about changes in aggregate demand and labour market conditions.

This story applies equally well whether firms set offering wages and workers search for employment or workers set asking wages and firms search for workers. A crucial ingredient of the search process, however, must be a degree of nonhomogeneity of both workers and jobs. Workers are all slightly different and are thus worth slightly different amounts to firms that know about them. Jobs and employers are also slightly different and yield different amounts of non-pecuniary benefits to workers. Negotiation between buyer and seller is therefore a necessary ingredient of labour market equilibrium.

The search theory goes part of the way in explaining both normal frictional unemployment and the positive correlation of the unemployment rate with the business cycle. And it is more realistic than the auction theory in that it allows for the fact that workers are actively searching for jobs rather than simply making employment decisions at existing market wage rates. Like the auction theory, however, it implies that quit rates should increase in recessions and decrease in booms, as workers will tend to regard their current employment situations as firm specific—when their firm lowers wages in a recession they will quit and begin a search for employment elsewhere at wage rates they expect not to have fallen. In fact, quit rates decrease during recessions and increase during booms, the opposite of what the search theory predicts. The search theory also fails to explain another very important fact—that firms actually lay people off during recessions and refuse to hire workers who are clearly willing to work for them at wages below what they are currently paying. Firms do not cut wages and maintain employment during periods of slack demand—instead, they tend to maintain wages and reduce the number of workers employed.

1.4 A Contract Theory

The third theory of price adjustment attempts to explain why quit rates decline and firms choose to lay off workers rather than reduce wages in recessions.[5] It begins by noting that individuals cannot diversify their human capital as easily as their non-human assets. Non-human wealth can be spread widely among bonds, equities, real estate, etc., while many individuals' human skills have only one avenue of employment. Thus, if workers are risk averse they will seek ways of insuring themselves against fluctuations in their incomes arising from variations in the demand for the narrow range of labour services they provide. The owners of firms, on the other hand, can diversify by owning little pieces of a large number of firms together with a variety of other assets. To the extent that the individual firm assumes some of the risk associated with fluctuations in the demand for its workers' human capital, its owners can diversify that risk away. It is profitable, therefore, for the firm to assume some of that risk in return for the acceptance by workers of a lower average level of wages. And it is profitable for workers to accept lower wages if the stability of their incomes can be increased. Firms and workers thus make a contract according to which workers will accept a lower wage in return for a guarantee of long-term income stability. This theory of price adjustment is thus referred to as the contract theory.

The contract between the firm and its workers may be an explicit one, hammered out in union-management negotiations, or it may be implicit, guaranteed solely by the fact that the firm must maintain its reputation as a 'good employer' if it is to be able to successfully hire workers over the long run. The essence of these contracts, whether explicit or implicit, is that

[5]For a more thorough and rigorous presentation of the argument developed here see Donald F. Gordon, "A Neo-Classical Theory of Keynesian Unemployment," *Economic Enquiry*, Vol. 12, 1974, 431–459.

the firm guarantees employment for a large fraction of its employees at real wage rates that reflect their mean or average marginal productivities over periods that may extend as long as a lifetime. In the most extreme cases the current wage paid could be viewed as one of many instalment payments in a lifetime contract. As a result of these considerations, the wage rate paid at any particular point in time may be above or below the workers' marginal products at that point in time.

The firm can be thought of as having several classes of 'tenured' employees. The lowest class gets layed off first when the demand for the firm's output declines, the next lowest class gets layed off next, and so forth, with the very senior employees getting layed off only if demand declines to the point where the survival of the firm is in jeopardy. Wages are maintained in the face of these layoffs and employment is kept at a point where the marginal product of labour is substantially below the wage rate the firm is paying its employed workers. Similarly, employment will expand in times when demand is high but the marginal product of labour will remain above the wage rate being paid.

This is illustrated in Figure 4. The individual firm's value marginal product, or marginal revenue product if it has monopoly power in product markets, is given by the vertical distance of the curve $\mathrm{V_0V_0}$ from the horizontal axis—this curve can be thought of as the firm's demand curve for labour. A temporary fall in demand in the industry shifts this curve down to $\mathrm{V_1V_1}$. To maximize its current-period profits at the wage rate W_0, the firm would cut employment to Q_1. Alternatively, workers would have to reduce their money wage demands to W_1 to induce the firm to continue to employ the quantity of labour Q_0 under the new demand conditions. Under its implicit contract with the workers, the firm will maintain the wage rate at W_0 and reduce employment to, say, Q_2. At this employment level the firm's current period profits will not be maximized, but its losses will not be as great as would occur if employment were maintained at Q_0. From the workers' point of view, there will be less layoffs than at Q_1 and no reduction in wages. This sacrifice of apparent current-period profits by the firm represents an absorption of some of the current-period losses that would otherwise be borne by the workers. It pays the firm to absorb these losses because the wage rate W_0 is lower than workers would require were the firm not prepared to absorb these temporary losses and risk of wage cuts or unemployment were therefore greater. By paying W_0 instead of a higher wage rate over the long-run, but taking less profit than could be obtained at that wage rate in the short run, the firm can achieve greater long-run profits. Similarly, when demand in the industry is temporarily high, and the value

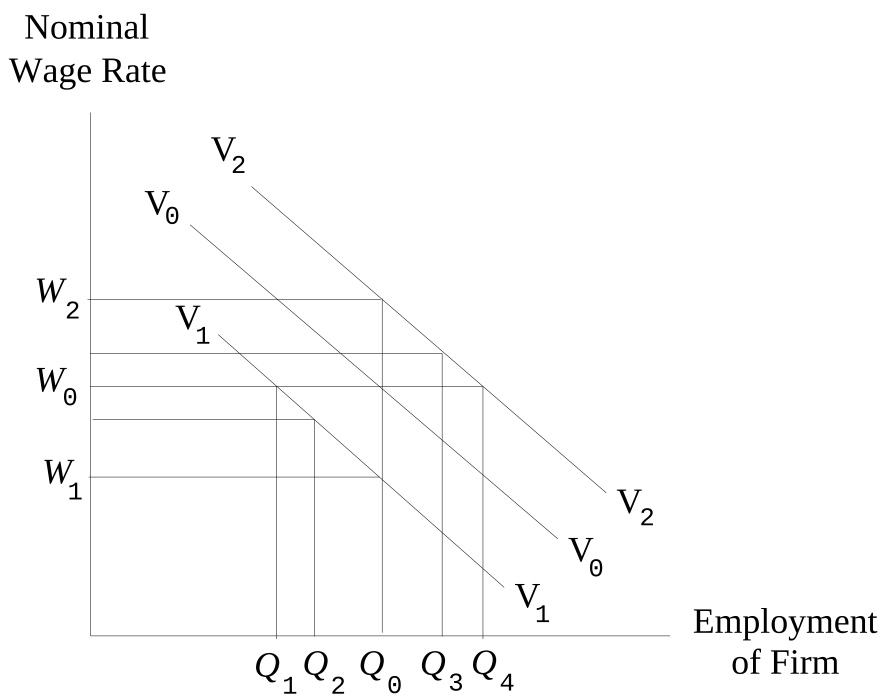

Figure 4: A Search Model of Wages and Employment.

marginal product of labour curve is V_2V_2, firms will increase employment to Q_3 and workers will not take advantage of the situation by demanding higher wages. Workers will accept wages below their value marginal products because they view the difference as a compensation to the firm for paying wages above the value marginal product in slack periods. Layed-off workers will return to the firm for wages less than could possibly be obtained elsewhere during the boom period because by adding to their seniority they can achieve a greater degree of tenure, thereby reducing their probability of being laid off during the next recession—at another firm they would have to start at the bottom of the seniority ladder.

Employment will not expand all the way to Q_4 in boom periods because the firm would have to hire workers that have some degree of tenure elsewhere, which it could not do without paying a wage rate above W_0.

The contract theory thus explains why quit rates are low in recessions and high in boom periods. And it also explains why firms, in slack periods, do not hire workers that apply to it willing to work at wage rates currently below those being paid.

While wages are set under explicit or implicit contractual arrangements—with a time horizon that may be several years long, they nevertheless are affected by current market conditions. Wages will be routinely increased each year by the expected rate of inflation, and adjusted in response to information about the level of aggregate demand to maintain real wages at the contractually agreed upon level. When, due to misinformation about market conditions, wages are set too high in relation to product demand, firms will end up hiring too few workers and producing too little output. When workers and firms eventually realize that they are pricing labour too high and firms are unable to produce average output levels at a cost consistent with market demand, the rates of increase of wages and perhaps even the level of wages will be lowered. And when wages are set too low in relation to product demand, firms will find themselves using too many workers on average and producing in excess of the normal full employment output level. Steps will be taken to bring wages up to a level consistent with the productivity of labour at normal average output levels. It follows that when aggregate demand unexpectedly increases as a result either of changes in the government's monetary and other policies or exogenous shocks arising from factors beyond anyone's control, firms will employ too many workers and the unemployment rate will fall below its normal or natural level. And when aggregate demand unexpectedly declines, firms will lay workers off and unemployment will rise above the natural rate.

1.5 Some Implications

Elements of all three of the theories of price adjustment outlined above have a role to play in explaining why money wages do not adjust immediately to shifts in aggregate demand so as to maintain employment at its long-run equilibrium level. And since the situation differs from industry to industry, a variety of stories based on the theoretical insights of the three approaches can be used to 'rationalize' particular events. However, all of the theories have one important thing in common. They predict that the unemployment rate will be at its natural level when price setters, be they workers or firms or both in any particular instance, are well informed about the current state of aggregate demand in the economy and future changes in it. And they all predict that unanticipated increases in aggregate demand will lead to a deviation of the unemployment rate below the natural rate and unanticipated decreases in aggregate demand will lead to greater than normal unemployment. Thus, if the government expands the money supply, the equilibrium money wage will rise. But if workers and firms are not aware of this shift in nominal aggregate demand they will continue to set wages at the old level. Firms will find it desirable to increase their labour force beyond the level originally planned upon and workers will find jobs a lot quicker than they had anticipated. The unemployment rate will fall below the natural level. Similarly, an unanticipated cut-back in the money supply will cause firms to employ less workers than they had originally planned and workers will find it taking them longer to find jobs than they had anticipated. The unemployment rate will rise above the natural level.

These deviations of the unemployment rate from the normal or natural level arise from an optimal response of price setters, both workers and firms, in the light of the information available to them. Non-optimality arises from the point of view of the economy as a whole only to the extent that price setters do not have the correct information about the state of the economy and this information could be costlessly distributed to them by the government. Changes in aggregate demand that are unanticipated both by government and the private sector will lead to deviations of the level of employment from the natural rate that will be the result of optimal decision making on the basis of all available information in the economy. Such deviations can hardly be regarded as resulting from non-optimal behaviour, since no one could have predicted the changes in aggregate demand that bring them about.

Where the government does not completely or convincingly inform the private sector about its policy plans, deviations of employment from the natural rate become a mechanism by which price setters learn that changes in government policy have taken place. When the government announces that it is fighting inflation by reducing the rate of monetary growth, and the private sector believes it, wages (and prices) will adjust immediately to the new aggregate demand conditions, and no deviations of unemployment from the natural rate will occur. However, such announcements are rarely believable. Politicians, to obtain political favour or votes, often make announcements that leave the impression that a particular policy will be followed but which can be interpreted in a variety of ways should the policy not be forthcoming as planned. Moreover, to the extent that a tightening of the money supply produces some unemployment and political stress, the policy makers may lose their nerve and rescind the policy before it has any substantial effect on the inflation rate. Price setters who acted on the basis of the announcement will be worse off than had they ignored it. Thus, where policy makers cannot believably inform the private sector of their plans, or where exogenous changes unrelated to government policy occur, a deviation of unemployment from its natural rate is necessary to inform wage and price setters that things have changed. High unemployment rates in recessions perform the function of informing wage setters that aggregate demand is growing less rapidly than they thought, and low unemployment rates in booms transmit to wage setters the information that aggregate demand is expanding at a more rapid rate than they had anticipated.

This analysis leads to a number of conclusions about the process of price adjustment. First, the natural rate of unemployment is not a 'full employment ceiling' but rather a level that will be achieved under conditions where price setters are well informed about the state of aggregate demand. Deviations below the natural rate are every bit as important as deviations above it.

Second, when we speak of wage rigidity or stickiness we should be referring to rigidity of the rate of growth of wages rather than their level. If prices have been inflating at 10 percent per year for the past five years, and everyone expects inflation to continue at that rate, money wages will be 'rigid' or 'sticky' with reference to that 10 percent growth path. A shift of the monetary 'regime' to result in a 5 percent inflation rate will result in a higher than normal unemployment rate until price setters catch on to what is happening. For a while, wages will be growing at the old rate while aggregate demand growth calls for wage increases that are 5 percent per year lower. Labour will be inadvertently priced out of the market. Eventually, as price setters learn about the new equilibrium inflation rate, wage increases will flatten out and eventually stabilize at an appropriate new growth rate. As this happens, unemployment will fall back to its natural level.

Third, the simultaneous existence of both inflation and unemployment is now quite possible. Unemployment will result when the actual rate of inflation is below firms' and workers' expected inflation rate.

Fourth, since unanticipated variations in the aggregate demand for output result in price changes in the same direction and changes in unemployment rates in the opposite direction, we can expect to observe a negative relationship between the rate of inflation and the rate of unemployment as long as the aggregate demand variations remain unanticipated. In other words there is good reason to expect a Phillips curve to appear when unanticipated shifts in aggregate demand are taking place. However, this Phillips curve is conditional upon a given expected rate of inflation. When a fully anticipated shift in aggregate demand occurs, the inflation rate will change without a change in the unemployment rate, and the Phillips curve will shift vertically. This is shown in Figure 5. The curve P_0P_0 gives the em-

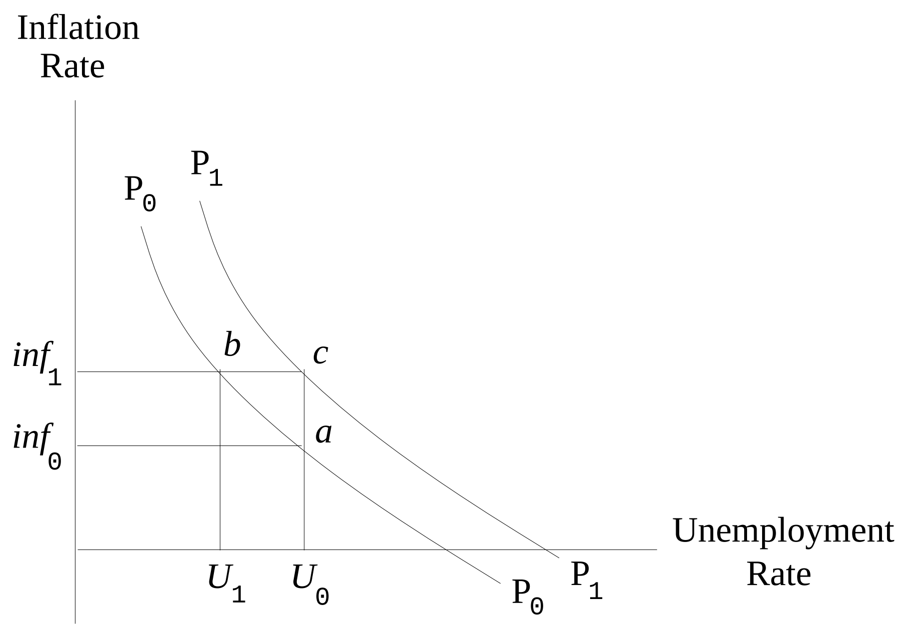

Figure 5: The Shifting Phillips Curve.

pirical relationship that will occur between the rate of inflation and rate of unemployment for given information and expectations possessed by the private sector. An unanticipated increase in aggregate demand will tend to drive the observed inflation-unemployment combination upward to the left along the curve from, say, *a* to *b*. Once it becomes known, however, that aggregate demand has shifted, the curve P_0P_0 will shift upward to P_1P_1 and the inflation-unemployment combination will move to *c*. There will be no permanent effect on the level of employment. Thus, policy makers do not face a tradeoff between inflation and unemployment in the long run. Moreover, if the increase in aggregate demand is fully anticipated at the time it occurs, the Phillips curve will shift immediately to P_1P_1, the inflation-unemployment combination will move immediately from *a* to *c*, and there will be no tradeoff in the short run either.[6]

[6] A very useful discussion of the Phillips Curve can be found in Olivier Blanchard, *Macroeconomics*, Prentice hall, 1997, Chapter 17.

1.6 Exercises

True or False: Explain your answer briefly.

a) Standard neo-classical microeconomic theory cannot explain the process by which prices adjust.

b) The demand for labour depends on actual wages while the supply of labour depends on expected wages.

c) Inflation always shifts the demand and supply curves for labour upward in proportion.

d) An anticipated increase in aggregate demand that does not occur will reduce the unemployment rate.

e) Unemployment arose in the Great Depression because workers, faced with declining wages, decided to take a holiday.

f) Firms that enter into implicit or explicit contracts with their workers regarding wages and working conditions are sacrificing short-run profits in order to increase long-run profits.

g) The contract theory can explain features of observed unemployment that auction and search theories cannot.

h) The better the unemployment insurance scheme a country has, the higher will be its natural rate of unemployment.

i) Wage 'rigidity' or 'stickiness' is entirely due to choices made by workers.

j) The existence of unemployment is clear evidence that people, both individually and collectively, are not pursuing utility maximizing policies.

k) In the long run the Phillips curve is vertical.

ESSAY 2

The Consumption Function

2.1 The Fisherian Analysis

The modern analysis of the determinants of consumption begins with the work of Irving Fisher.[1] Consider a consumer who lives only two periods, receives incomes Y_j in the jth period (j = 0,1), and can borrow and lend at the real interest rate r. Utility depends entirely on consumption in the two periods.

$$U = U(C_0, C_1) \tag{2.1}$$

The consumer's problem is to choose the consumption levels C_0 and C_1 so as to maximize utility. The analysis is presented in Figure 1. If the consumer wants to spend her entire lifetime income on first period consumption, she can spend that year's income in its entirety plus the maximum amount that can be borrowed against next year's income.

$$m_0 = Y_0 + \frac{Y_1}{1+r} \tag{2.2}$$

Alternatively, the consumer's maximum possible consumption in year 1 is

$$m_1 = Y_0(1+r) + Y_1 \tag{2.3}$$

which is the amount can be accumulated in year 1 by saving Y_0 plus the amount of income received in year 1. The rate at which the consumer can

[1] Irving Fisher, *The Theory of Interest*, New York: Macmillan, 1930, Chapter 10.

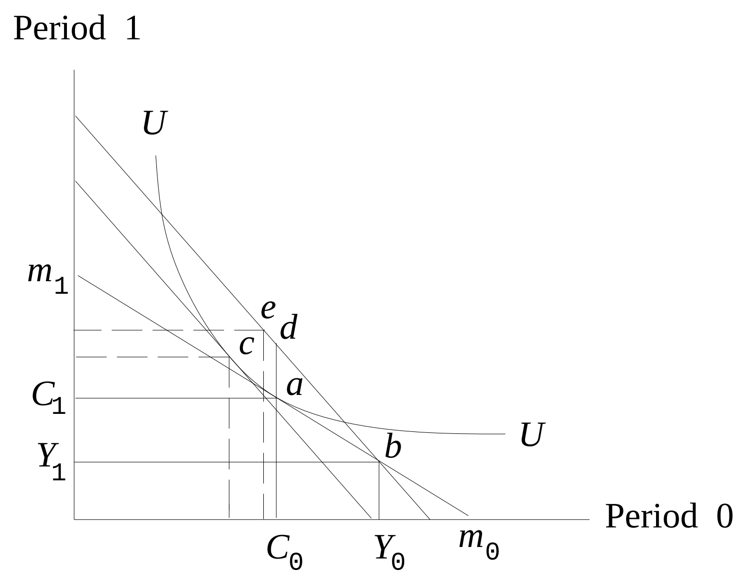

Figure 1

substitute consumption in year 0 for consumption in year 1, which is the slope of the line $m_1 m_0$ in Figure 1, is

$$\frac{dC_1}{dC_0} = -\frac{m_1}{m_0} = -(1+r) \tag{2.4}$$

Utility maximization will put the consumer on the highest achievable indifference curve. Along this curve utility will be constant, implying that at any point such as a in Figure 1

$$dU = \frac{\partial U}{\partial C_0} dC_0 + \frac{\partial U}{\partial C_1} dC_1 = 0. \tag{2.5}$$

The slope of the consumer's indifference curve at point a is thus

$$\frac{dC_1}{dC_0} = -\frac{\partial U/\partial C_0}{\partial U/\partial C_1}. \tag{2.6}$$

In equilibrium

$$\frac{\partial U/\partial C_0}{\partial U/\partial C_1} = -(1+r). \tag{2.7}$$

On the left side of the equality is the marginal rate at which the consumer is willing to substitute present and future consumption and on the right side is the rate at which the consumer is able to exchange present and future consumption. In equilibrium the desired rate must equal the possible rate—otherwise the consumer will have an incentive to reallocate consumption between periods.

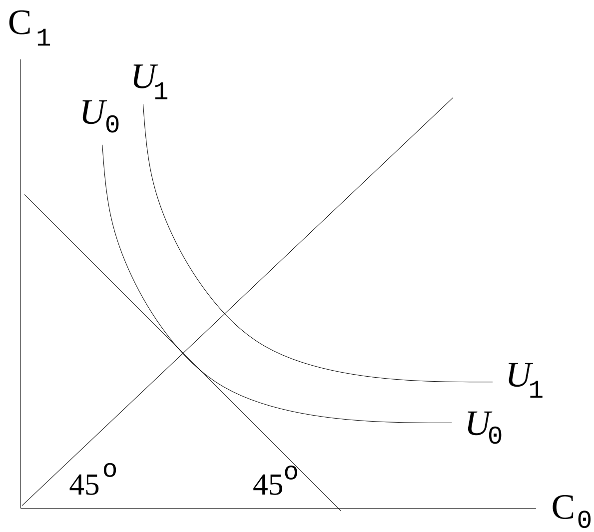

Figure 2

It is sometimes useful to write the utility function as the sum of separate and identical utility functions for each period (called felicity functions), with future utility discounted back to the present in accordance with the individual's time preference.

$$U = U(C_0) + \frac{U(C_1)}{1+\rho} \tag{2.8}$$

where ρ is called the *rate of time preference*. The meaning of ρ can be seen more clearly with reference to Figure 2. In that figure, the indifference curves are *homothetic*—that is, symmetrical around a 45^o line through the origin. And these indifference curves all have slopes equal to -1 at the point they cross the 45^o line. The rate of time preference ρ is equal to the slope of the indifference curve in absolute value minus unity at the point where it crosses the 45^o line. In Figure 2, the utility function is such that $\rho = 0$

—i.e., the slopes of all the indifference curves where they cross the 45^o line is -1. In that case, we say that the individual has zero time preference. Confronted with a zero rate of interest—that is, an intertemporal budget constraint with slope equal to minus one—the individual will have no greater preference at the margin for goods this period rather than next period, and will therefore consume the same amount in both periods. If the indifference curves have a slope steeper than -1 at the point where they cross the 45^o line through the origin, the individual will have a positive value of ρ and, given a zero rate of interest indicated by the budget line *ab* in Figure 3, will consume more in period 0 than in period 1. Similarly, when the indifference curves have a slope flatter than minus one at the point where they cross the 45^o line through the origin, ρ will be negative. The individual will have a negative time preference and consume less in period 0 than in period 1 when the interest rate is zero. If there is a positive rate of interest equal to the rate of time preference—that is, $\rho = r$ —the slope of the consumer's budget line along the 45^o line through the origin will equal the slope of her indifference curve, and the same amount will be consumed in each period. This is indicated by the budget line *cd* in Figure 3. The marginal rate of substitution at the equilibrium point, given by the slope of the indifference curve, equals

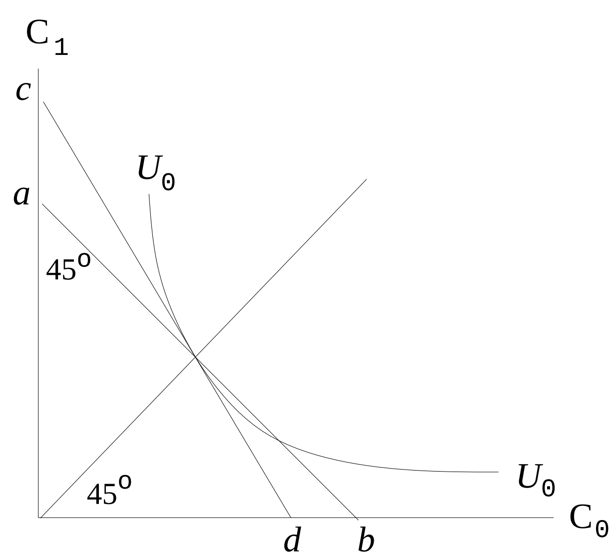

Figure 3

$$\frac{dC_1}{dC_0} = -\frac{\partial U(C_0)/\partial C_0}{\partial U(C_1)/\partial C_1}\,(1+\rho) = -\frac{U'(C_0)}{U'(C_1)}\,(1+\rho) \tag{2.9}$$

and the individual's equilibrium condition now becomes

$$\frac{U'(C_0)}{U'(C_1)}\,(1+\rho) = (1+r). \tag{2.10}$$

If the rate of time preference equals the rate of interest, consumption will be allocated between the periods so that $U'(C_0) = U'(C_1)$. The level of consumption will therefore have to be the same in both periods. If the rate of time preference exceeds the rate of interest, $U'(C_0)$ will be less than $U'(C_1)$ and, since there is diminishing marginal utility in each period, C_0 will necessarily exceed C_1.[2]

Two crucial results follow from this analysis. First, the levels of consumption in both periods depend on the interest rate and on incomes in both periods—current income affects current consumption only to the extent that it affects overall wealth. Second, a change in the interest rate will have both wealth and substitution effects on consumption if the initial level of savings is non-zero.

To expand on the latter point, suppose that starting from an initial two-period income combination at point b in Figure 1 the rate of interest rises. The budget constraint rotates clockwise around the point b, putting the consumer on a higher indifference curve. Equilibrium consumption in the first year will increase or decrease depending upon whether the individual's new indifference curve is tangent to the right or to the left of point d. The movement to this new equilibrium, which occurs at a point such as e, will be composed of a substitution effect, represented by the movement from a to c along the consumer's original indifference curve, and a wealth effect, represented by the movement from c on the original indifference curve to e on the new one.

In the many period-case the consumer, looking at the situation at time t, maximizes

$$U(C_t, C_{t+1}, C_{t+2}, C_{t+3}, \ldots\ldots C_{t+T}), \tag{2.11}$$

which in the temporally separable case becomes

[2]Note that when the utility function is given by (2.8) all indifference curves must have the same slope along the 45^o line extending outward from the origin. Since $C_0 = C_1$ and ρ is constant everywhere along that line, the marginal utilities of consumption in the two periods as well as their ratio must also be constant along the line.

$$U(C_t + \frac{C_{t+1}}{1+\rho} + \frac{C_{t+2}}{(1+\rho)^2} + \frac{C_{t+3}}{(1+\rho)^3} + \ldots\ldots + \frac{C_{t+T}}{(1+\rho)^T} \tag{2.12}$$

The consumer maximizes her utility function subject to the constraint that the present value of consumption must equal the present value of income—that is,

$$\begin{aligned} C_0 + \frac{C_1}{1+r} + \frac{C_2}{(1+r)^2} + \ldots\ldots + \frac{C_T}{(1+r)^T} & \\ = \quad Y_0 + \frac{Y_1}{1+r} + \frac{Y_2}{(1+r)^2} + \ldots\ldots + \frac{Y_T}{(1+r)^T} & \end{aligned} \tag{2.13}$$

The problem with temporal separability is that it assumes that the substitutability of consumption between two periods is independent of the level of consumption in all other periods. This would imply, for example, that one's marginal rate of substitution between lunch and dinner is independent of the amount one had for breakfast.

The maximization process results in an equilibrium level of consumption in year t equal to

$$C_t = C_t(Y_t, Y_{t+1}, Y_{t+2}, Y_{t+3}, \ldots\ldots Y_{t+T}, r). \tag{2.14}$$

The nature of the individual's preferences between present and future consumption is contained in the form of the function $C_t(\ldots)$. In the time separable case, equation (2.14) becomes

$$C_t = C_t(Y_t, Y_{t+1}, Y_{t+2}, Y_{t+3}, \ldots\ldots Y_{t+T}, \rho, r). \tag{2.15}$$

The individual's rate of preference is now an argument in $C_t(\ldots)$ rather than embedded in the form of that function. And the specification of time separability imposes restrictions on the functional form—namely, that the indifference curves all have the same slope on every ray extending outward from the origin.

Equations (2.14) and (2.15) tell us only that consumption in any period depends on the incomes received in it and all subsequent periods, on time preference, and on the rate of interest at which the individual can borrow and lend. Apart from constancy of time preference in the case of (2.15), they tell us nothing about the nature of the relationship between future incomes and current consumption. Moreover, the actual allocation of consumption through time depends not only on preferences but on the individual's information about future incomes. To apply this theory to real problems, we must impose some hypotheses about how individuals acquire and react to information about their future levels of income. Efforts to do this, which are essential before the theory can be confronted with empirical evidence, must also recognize the fact that the age distribution of the population will have an effect on aggregate consumption, given that the time preference of the old and the young may be different.

2.2 Specifying the Consumption Function for Empirical Analysis

Keynesian theory has tended to view consumption as dependent on current disposable income with little if any acknowledgment of possible effects of changes in future income. Numerous empirical studies have established that consumption tends to increase less than proportionally with increases in current income, and that the ratio of consumption to current income is therefore negatively related to the latter variable, over the business cycle and across families at any given point in time. In other words the marginal propensity to consume out of changes in current income has been shown to be less than the average propensity to consume with, of course, the average propensity to consume being less than unity. An example is the following estimate of the U.S. consumption function for the period 1929–40,

$$\begin{array}{ccccccc} C & = & \underset{(.126)}{1.326} & + & \underset{(.025)}{0.429} & Y_D \end{array}$$

where C and Y_D are consumption and real disposable income, respectively.[3] The numbers in parentheses are the standard deviations of the coefficient estimates. The marginal propensity to consume is .429, and the fact that the constant term is positive with P-value equal to .000000498 clearly suggests that the marginal propensity to consume is less than the average propensity to consume at every level of income. The P-value is the probability of observing a more extreme value for the coefficient than the one observed if the true value of the coefficient were zero.[4] Indeed, as is clear from Figure 4, the average propensity to consume, which is the slope of the ray from the

[3] The Data are from U.S. Department of Commerce, *National Income and Product Accounts of the United States, 1929–88.*

[4] The P-value is obtained by calculating a t-statistic, which is equal to the ratio of the estimated value of the coefficient to its standard error, and then obtaining the probability that t will exceed that value from a set of statistical tables. Such tables can be found at the back of any textbook in statistics.

origin to the consumption-income line YC, decreases as income rises and is always greater than the marginal propensity to consume, which is given by the slope of the YC line. This would not be true if the YC line passed through the origin and the constant term in the regression was therefore zero—the marginal and average propensities to consume would be the same.

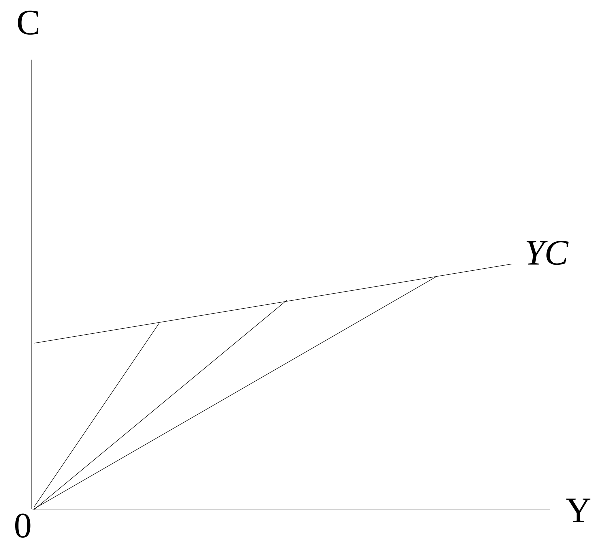

Figure 4: Consumption (vertical axis) vs. Income (horizontal axis).

One apparent implication of the average propensity to consume being lower when income is higher is that as the economy grows through time the rich will tend to get richer and the poor poorer because the rich will be saving higher fractions of their incomes than will the poor.

When the relationship between consumption and income is examined over a very long period of time, however, it is clear that the ratio of consumption to income has not tended to decline as society got richer. Figure 5 shows the relationship between consumption per capita and real income per capita for the United States over the 70-year period 1929–99. When we fit a straight line to these data we obtain

$$\begin{array}{ccccccc} C & = & -.37 & + & 0.655 & Y_D \\ & & (.18) & & (.0152) & \end{array}$$

which yields a negative constant term suggesting that, if anything, the average propensity to consume has tended to rise with income.

For twenty years after the Second World War, major research efforts were devoted to attempts to reconcile in terms of economic theory the conflict between the consumption function observed over short periods and for individual groups of consumers in any given period and the consumption function observed over long periods such as the one obtained from the data in Figure 5. The crucial principle involved in this reconciliation utilizes the fact that, as the Fisherian analysis shows, consumers are forward-looking. They base their consumption not just on current income, but on the future incomes expected over their lifetimes. The view of consumer behaviour that has emerged can be termed the *life-cycle permanent income* approach, which has two branches—the permanent income theory and the life-cycle theory.

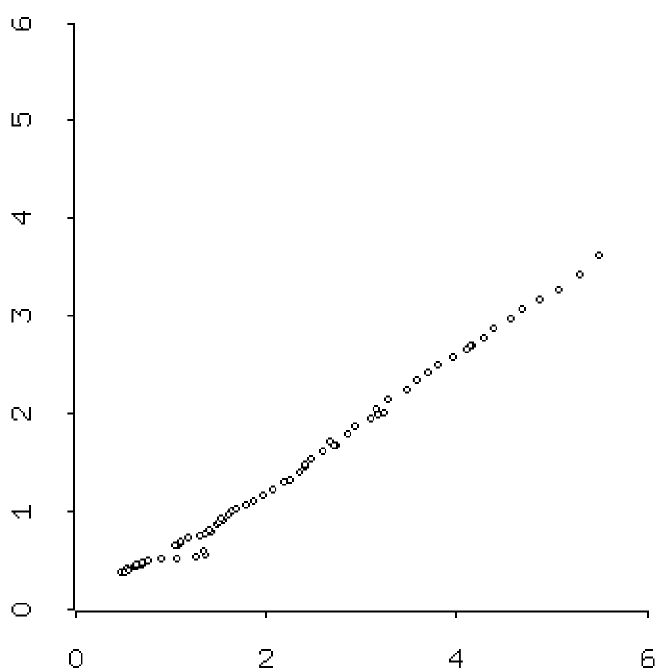

Figure 5: Personal Consumption per capita (vertical axis) vs. real GNP per capita (horizontal axis) for the United States, 1929–99. The scale is thousands of 1982 dollars. Source: U.S. Department of Commerce, *National Income and Product Accounts of the United States, 1929–88*, and *Statistical Abstract of the United States, 2001.*

2.2.1 The Permanent Income Approach

Milton Friedman developed and empirically applied the argument that consumption in any period depends on consumers' expected or permanent income flow rather than the current level of income, which depends on a variety of transitory factors.[5] He provided no clear-cut definition of permanent income, although a good definition might be the maximum constant consumption flow that could be sustained in perpetuity, given the individual's current resource endowment. This would equal

$$Y_P = rP_{V0} = r\left[Y_0 + \frac{Y_1}{1+r} + \frac{Y_2}{(1+r)^2} + \ldots\ldots + \frac{Y_T}{(1+r)^T}\right] \qquad (2.16)$$

where P_{V0} is the present value of income in the year zero, a measure of wealth as viewed from that period.

The problem with this definition is that the individual will not consume a constant amount over the rest of his lifetime, but will probably consume more than his permanent income when he is very young and has limited resources, then consume substantially less than his permanent income in middle life in order to build up a capital stock which will be drawn down in his retirement years when consumption will again exceed permanent income. Nevertheless, defining income in this way focuses on the effects of the variability of income on the relationship between consumption and income.

After dividing current income into its permanent and transitory components, Friedman postulated a) that transitory and permanent income are uncorrelated, and b) that consumption is a constant fraction k of permanent income and is uncorrelated with transitory income. He then argued that a significant proportion of the variance of income over the business cycle is due to the transitory components, while over many generations transitory income variations make up a very small proportion of the total variance. Hence, the marginal propensity to consume estimated from time series for short periods of a decade or so will be lower than the marginal propensity to consume estimated from time series covering many decades. Similarly, the marginal propensities to consume estimated from cross-sectional data for groups whose incomes are highly variable (for example, farmers and fishermen) should be lower than the marginal propensities to consume for groups whose incomes are highly stable (for example, civil servants).

The permanent income theory is shown diagrammatically in Figure 6. Individuals whose income in a particular period equals the average for the

[5] See Milton Friedman, *A Theory of the Consumption Function*, University of Chicago Press, 1957.

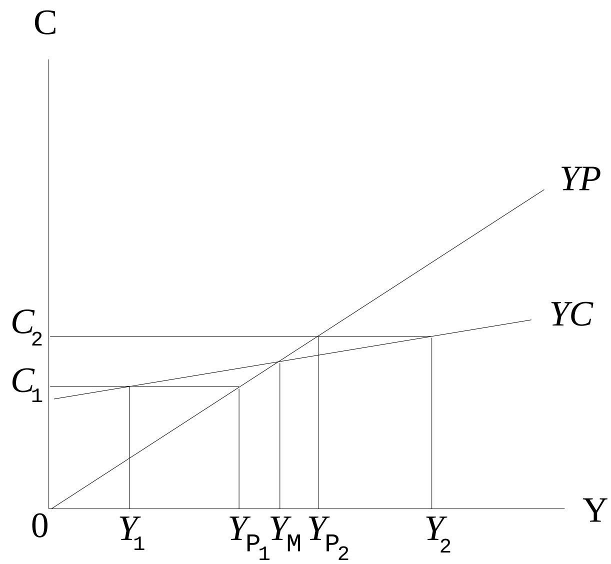

Figure 6: Consumption (vertical axis) vs. Income (horizontal axis).

group, shown as Y_M, will on average have zero transitory income. Their consumption will be on the permanent income consumption function kY_p, shown as the line YP, and will on average equal the average consumption for the group. Individuals whose current income is low, say at Y_1, will tend to have lower permanent incomes than the average for the group, but will also typically have negative transitory income as well. These individuals' average permanent income is shown on as Y_{P1} and their average consumption is C_1. Individuals whose current income is high, averaging Y_2, will also typically have high average levels of permanent income, say Y_{P2}, and their average consumption is C_2. The relationship between consumption and current income is thus given in Figure 6 by the line YC, which represents the current income consumption function.

In testing the permanent income hypothesis, Friedman set permanent income equal to the average income of the group in cross-sectional studies. In time-series studies he proposed that the consumer at each point in time estimates his permanent income on the basis of past experience. Permanent income is adjusted from period to period by some fraction λ of the difference between last period's current income and last period's permanent income.

$$Y_{Pt} - Y_{P(t-1)} = \lambda[Y_{t-1} - Y_{P(t-1)}] \tag{2.17}$$

This expression can be rearranged to yield

$$Y_{Pt} = \lambda\, Y_{t-1} - (1-\lambda)\, Y_{P(t-1)} \tag{2.18}$$

which can be used to calculate permanent income on the basis of a chosen value of λ and some estimate of permanent income in some initial period (usually current income in that period). By repeated substitution, equation (2.18) can be reduced to

$$\begin{aligned} Y_{Pt} \;=\;& \lambda\, Y_{t-1} + (1-\lambda)\,\lambda\, Y_{t-2} + (1-\lambda)^2 \lambda\, Y_{t-3} + (1-\lambda)^3 \lambda\, Y_{t-4} + \ldots\ldots \\ & \ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{2.19}$$

which expresses permanent income as a weighted average of past incomes with the weights declining geometrically as we move back into the past. For example, if we select $\lambda = .4$, the weights will be .4, .21, .144, .0864, etc. The sum of the weights is always unity.[6] One way of finding the appropriate value of λ in any particular situation is to regress consumption on permanent income as defined by equation (2.19) for different values of λ, choosing the value which yields the best fit to the data.

The problem with the permanent income approach is in defining permanent income. Is permanent income simply the expected level of income at each point in time? Or is it some average of lifetime income? Why should permanent income be best determined by past income levels when information about the future not available in the past may now be available? In other words, how is expected future income affected by past incomes as opposed to current news about future incomes?

2.2.2 The Life Cycle Theory

The life-cycle theory, developed by Franco Modigliani and several collaborators,[7] takes into consideration changes in the pattern of consumption in relation to income over the individual's life cycle. Individuals typically borrow against future income to maintain consumption above current income in the early years of their lives, pay off these debts and accumulate assets

[6]This can be seen by noting that the sum of the weights equals

$$\lambda[1 + (1-\lambda) + (1-\lambda)^2 + (1-\lambda)^3 + \ldots\ldots\ldots] = \lambda\,\frac{1}{1-(1-\lambda)} = 1.$$

[7]The seminal paper is Franco Modigliani and Richard Blumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data", in K. K. Kurihara, ed., *Post-Keynesian Economics*, Rutgers University Press, 1954, 388–436.

by saving out of their current income (and thereby consuming less than it) in the middle years, and then draw down these assets by consuming more than their current income during retirement.

Suppose for the sake of argument that there are no transitory variations in income. Then an individual starting out with given skills at the beginning of his life would have a consumption and current income pattern like that in Figure 7. In any population of households, those whose income is high relative to consumption will tend to be middle aged, while those whose consumption is high in relation to income will tend to be the old and the young. A regression of consumption on income will yield a marginal propensity to consume less than the average propensity to consume. Yet as the income of the whole society rises, the levels of consumption and income of each age group will tend to rise in proportion. Thus a regression of average consumption of the population on the average income of the population as the income of the society rises over time will, in the absence of changes in the age distribution of the population, yield a marginal propensity to consume that is more or less equal to the average propensity to consume. The two curves in Figure 7 will shift upward in the same proportion. The basic stylized facts about consumption and income are thus explained by this theory independently of any transitory variations of income. But the life-cycle theory complements rather than contradicts the permanent income hypothesis—we would expect both theories to apply simultaneously.

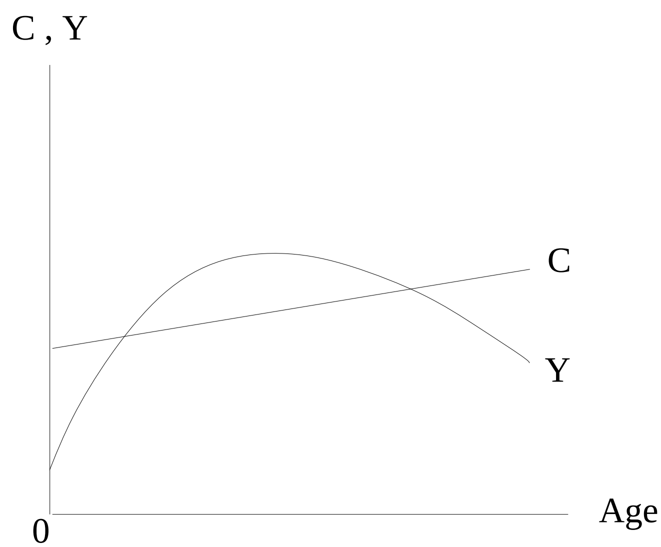

Figure 7: The Life-Cycle Hypthesis

2.3 Rational Expectations and Consumer Behaviour

The empirical applications of the permanent income and life-cycle hypotheses have been backward-looking in the sense that expectations of the future course of income depend either directly on current income or adaptively on incomes experienced in the past. For example, equation (2.17) postulates that the change in permanent income from last period to this one is a constant fraction of last period's difference between current and permanent income and, as a result, this period's permanent income is a weighted average of all past periods' current incomes with the weights declining exponentially as we go back into the past. There is no role for information about the future not contained in past incomes.

In fact, however, the consumer can be thought of as maximizing the expected present value of her discounted future utility flow subject to the current value of her non-human assets and an uncertain future flow of labour income. At each point in time t, the consumer knows only her current labour income and the current value of her assets. She must form an expectation or forecast of her future labour income and future changes in the value of her assets. This forecast is made on the basis of information available to her at time t.

Suppose that new information becomes available to the consumer at time t. Then the current and all expected future consumption levels will immediately adjust to take account of that new information. Moreover, all information contained in the current and past values of income will be reflected in the levels of current and expected future consumption. To the extent that next period's actual consumption differs from the level expected on the basis of present information, that change will be the result of new information available next period that was not available this period. All information available as of this year will be fully reflected in this year's consumption. One should therefore be able to estimate consumption in any given year as the result of two information sets—information available in the previous year, as reflected in the previous year's consumption, and new information acquired this year. The expected effect, as of last year, of new information available this year must be zero if consumers' expectations are formed rationally—if the new information were more likely to lead to an increase in this year's wealth and consumption than decrease it, the consumer's wealth calculations last year would have taken that probability into consideration and consumption last year would have been higher. So an appropriate equation estimating current consumption would be

$$C_t = \delta + \gamma\, C_{t-1} + \epsilon_t \tag{2.20}$$

where ϵ_t is stochastic with mean zero, representing the effect of new information available at time t. Since C_{t-1} reflects all information available at time $t-1$, no other variables such as income at time $t-1$ or earlier or consumption at time $t-2$ or earlier should have significant coefficients when introduced into the equation. This formulation was first put forward by Robert Hall.[8]

The magnitude of γ can be determined as follows.[9] At any point in time t, the utility maximizing consumer will have allocated his consumption so that the marginal utility of consumption now will equal the discounted value of the additional utility that can be expected next year by shifting that unit of consumption forward to period $t+1$. That is

$$U'(C_t)\Delta C_t = \frac{E_t\{U'(C_{t+1})\}}{1+\rho}\,[(1+r)\,\Delta C_t] \tag{2.21}$$

where ΔC_t is the 'small' amount of consumption given up this year, the term $E_t\{U'(C_{t-1})\}[(1+r)\,\Delta C_t]$ is the amount of utility expected to be obtained from the resulting increase in consumption next year, and the reciprocal of $(1+\rho)$ is the discount factor, which gives the present value in the current year of the additional utility expected next year. Cancelling out ΔC_t and rearranging the expression, we obtain

$$E_t\{U'(C_{t+1})\} = \frac{1+\rho}{1+r}\;U'(C_t) \tag{2.22}$$

This equation is known as the Euler equation. Suppose now that the utility function is approximately quadratic, so that the marginal utility of consumption can be represented as a linear function of the level of consumption—that is,

$$U'(C_t) = \alpha + U''C_t \tag{2.23}$$

where U'' is a constant equal to the derivative of $U'(C_t)$ with respect to C_t. This is shown in Figure 8 where $U'(C_t)$ is on the vertical axis and C_t is on the

[8]Robert E. Hall, "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence", *Journal of Political Economy*, Vol. 86, No. 6, 1978, 971–987, reprinted in Robert E. Lucas Jr. and Thomas J. Sargent, eds., *Rational Expectations and Economic Practice*, Vol. 2, University of Minnesota Press, 1981, 501–517.

[9]For a more detailed and rigorous development of the argument presented here, see Hall's paper referred to in the previous footnote.

horizontal one. The slope of the marginal utility of consumption line is U'', which is negative. Lagging equation (2.22) one period, substituting equation (2.23) into the left-hand side, and then substituting a one-period lag of (2.23) into the right-hand side, we obtain equation (2.20) with $\delta = \alpha\,(\gamma - 1)$ and $\gamma = (1+\rho)/(1+r)$. If consumption is growing through time, $\gamma > 1$ and $\delta < 0$; if it is constant through time, $\gamma = 1$ and $\delta = 0$.

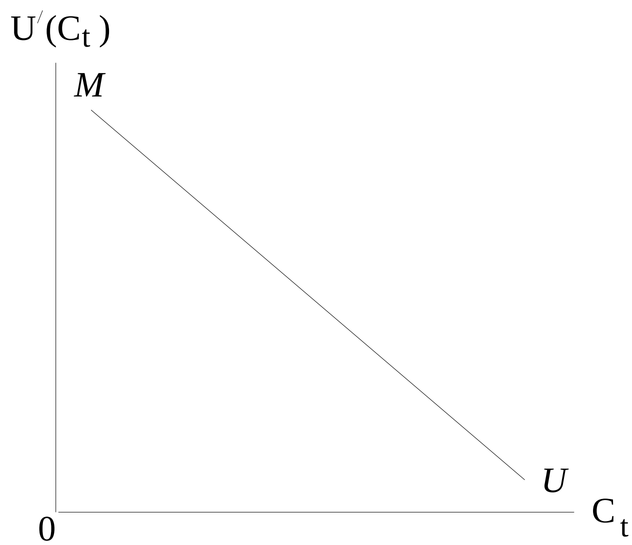

Figure 8: The Marginal Utility of Consumption

We can do a rough test of this theory with the U.S. annual data used previously. We run a regression of current consumption on last period's consumption plus four lags of real income. The resulting estimated equation is

$$\begin{array}{cccccccccc} C_t & = & -.0109 & + & 1.01288 & C_{(t-1)} & + & .0473 & Y_{(t-1)} \\ & & (.0522) & & (.0378) & & & (.0412) & \\ & & & - & .0315 & Y_{(t-2)} & - & .0098 & Y_{(t-3)} \\ & & & & (.0646) & & & (.0643) & \\ & & & & & & - & .0005 & Y_{(t-4)} \\ & & & & & & & (.0394) & \end{array}$$

where C and Y refer to the levels of per-capita real consumption and real income. The P-value for the F-statistic testing the null hypothesis that the coefficients of all the lagged real income terms are zero is .655 —this gives the probability of observing a value of the F-statistic as high or higher than the one observed (= .613) if the true coefficients of all the lagged real income terms are zero. This is consistent with the rational expectations hypothesis put forward by Hall and is consistent with similar estimating equations he obtained using quarterly U.S. per capita real income and consumption data for the period 1948-71. As a further test of whether lagged wealth had an effect on current consumption additional to lagged consumption Hall used, instead of income, four lags of Standard and Poor's comprehensive index of the prices of stocks deflated by the implicit deflator for consumption of non-durables and services and divided by population. The resulting estimated equation was

$$\begin{array}{ccccccccc} C_t & = & \underset{(8.0)}{-\,22} & + & \underset{(.004)}{1.012} & C_{(t-1)} & + & \underset{(.0533)}{.223} & S_{(t-1)} \\ & & & - & \underset{(.0832)}{.258} & S_{(t-2)} & + & \underset{(.083)}{.167} & S_{(t-3)} \\ & & & & & & - & \underset{(.0511)}{.120} & S_{(t-4)} \end{array}$$

where S refers to the stock-price variable. Each lagged stock price coefficient is significantly different from zero—the P-values for the coefficients of the four lags are, respectively, .000017, .0013, .0238, and .0106. The F-statistic for the null hypothesis that all coefficients are zero is 6.1, with a P-value of .00025. On the basis of this evidence, Hall rejects the hypothesis that consumption in year t cannot be predicted by any variable dated $t-1$ or earlier other than C_{t-1}. He interprets this not as a rejection of the hypothesis that consumption depends on permanent income, but as an indication that some part of consumption takes time to adjust to a change in permanent income. In this event, variables correlated with permanent income in $t-1$ will help in predicting the change in consumption in period t since part of that change is a lagged response to changes in permanent income. He regards the evidence against the typical Friedman distributed lag measure of permanent income as quite strong.

2.4 The Role of News vs. Past History in Consumption Decisions

Friedman measures permanent income as a weighted average of past incomes with the weights declining as we go back into the past. Hall argues that past income should have no effect on current consumption once last period's consumption has been taken into consideration. If last period's consumption was a constant fraction of permanent income then it is a good measure of last period's permanent income. This year's permanent income will differ from last year's only to the extent that there is new information, or 'news' about future income. Any information available last year about current and future incomes was already included in last year's permanent income and so cannot cause permanent income to change between last year and this year. Does this imply that Friedman's definition of permanent income is inappropriate?

Individuals can be expected to use both historical experience and current news in determining their permanent income. In a world where there is no news—where current information about the future is non-existent—people will forecast their future incomes on the basis of the behaviour of income in the past. Current income changes will modify the individual's expectations of future income on the basis of the relationships between past periods' current income and the evolution of income that subsequently took place. It will still be true, however, that last period's consumption will be based on all information available at that time and that current income will affect permanent income and current consumption to the extent that it improves the individual's understanding of the process by which income is evolving.

In fact, of course, news is very important. Current information may be received about future policies of the government, the future likely course of technological change, etc., which will lead individuals to change substantially their forecasts of future income. The evolution of income in the past may have some bearing on expectations about how the future will unfold, but forecasts of future income based solely on past incomes are not likely to be useful. Again, all information available last period is used in establishing this period's permanent income and consumption. Current period permanent income and consumption will differ from last period's permanent income and consumption on the basis of current information, some of which will be contained in movements of current income and some of which will be unrelated to the current innovation (though not to future innovations) of income. It is clear that only in special circumstances can a good estimate of permanent income be obtained from modelling past income alone. Friedman's distributed lag measure of permanent income has been useful for explaining consumption largely because, like 'true' permanent income, it has a smaller variance than current income.

2.5 Exercises

1. True or False: Explain your answer briefly.

   a) The Keynesian consumption function provides clear evident that through time the rich will get richer and the poor will get poorer.

   b) The ratio of consumption to income varies countercyclically.

   c) Cross-sectionally, the savings-income ratio increases as income increases. This implies that the marginal propensity to consume is greater than the average propensity to consume.

   d) In general, if an individual has zero time preference and the utility function
   $$U = U(C_0, C_1, C_2, \ldots\ldots, C_n)$$
   the absence of time preference implies that the marginal rate of substitution between consumption in year i and consumption in year j is zero.

   e) Zero time preference implies that consumption will be the same in all years regardless of income.

   f) The marginal propensity to consume out of permanent income equals the average propensity to consume, while the marginal and average propensities to consume out of transitory income are zero.

   f) The marginal propensity to save out of transitory income is unity.

   g) The marginal propensity to consume is less than the average propensity to consume because of errors in measuring permanent income.

2. Assume that the fraction of permanent income consumed is constant but that the marginal propensity to consume is less than the average propensity to consume in calculations from a set of cross-sectional data. Using a diagram, explain this result on the basis of Friedman's hypothesis.

3. Using the Ando-Modigliani life-cycle hypothesis, explain why the average fraction of income consumed of a cross-sectional group tends to be constant as it income grows while a regression of consumption on current income for the same group tends to have a marginal propensity to consume less than the average propensity to consume.

4. Explain the difference between the life cycle hypothesis of Ando and Modigiani and the permanent income hypothesis of Milton Friedman. Are the theories inconsistent with each other? Which of the two theories is best?

5. Consider a standard two-period diagrammatic representation of intertemporal choice. Measure the earlier year's consumption and income on the horizontal axis.

a) Suppose that the intercept of the budget line with the horizontal axis is 140, income in year 0 is 50, and the interest rate is 3%. Calculate what the level of income in year 1 must have been. Then calculate the intercept of the budget line with the vertical axis.

b) Prove that the slope of the budget line equals $-(1+r)$ where r is the rate of interest.

c) Suppose that $C_1 = 1.1\,C_0$. Calculate savings in year 0.

d) If the individual has zero time preference, will the slope of the indifference curve at a 45^o ray from the origin equal -1.0?

e) Demonstrate on your graph that an increase in current income will result in a less-than-proportional increase in current consumption.

f) How would one define the level of wealth on this graph?

6. Consider an individual consumer with the two-period utility function

$$U = C_1^{1/2} C_2^{1/2}$$

and suppose that income is 150 in year 1 and 50 in year 2.

a) Show that the individual has zero time preference.

b) Show that consumption will be the same in both years if the interest rate is zero regardless of the levels of income in the two years.

c) Calculate the budget constraint when the interest rate is 5%.

d) Using the budget constraint in c) and the utility function in a), show how you would calculate the levels of consumption in the two years and the level of savings in year 1.

e) If income in both years doubled, how would consumption in the two years be affected?

f) If income in year 1 increases to 200, would consumption increase by more than one-third its original level or less?

g) In case f), how much would consumption increase in year one if the interest rate were zero.

7. Consider a group of individuals whose mean income turns out to be 100, but who experience substantial fluctuations in current income. People consume 75 percent of their permanent income every year. In general, about 50 percent of any deviations of individuals' current income above or below the average for the group reflects differences in permanent income. Plot the current income consumption function.

Now suppose that improvements in technology in a subsequent period lead to an increase in everyone's income by 25 percent. Plot the current income consumption function for this period. Plot the permanent income consumption function. If one were to fit a current income consumption function in the two periods together, what would it look like in comparison to the consumption functions for the two periods separately and the permanent income consumption function?

ESSAY 3

The Investment Function

Elementary closed-economy analyses typically present a rudimentary analysis of investment behaviour of the sort illustrated in Figure 1. All potential projects in the economy are ranked from left to right in the figure according to the rate of return they yield. The highest return project yields r_m and the returns progressively fall as the level of investment expands and lower and lower return projects are resorted to. At a level of investment equal to I_m, all projects having positive return have been undertaken, and further investment will involve projects that have negative returns.

Profit maximizing firms will undertake all projects whose returns exceed the rate of interest at which they can borrow to finance them. If the real interest rate in the economy equals r_0, therefore, the level of investment will equal I_0. A fall in the real interest rate will thus lead to an expansion of investment to include projects that are profitable at the new rate of interest but were not profitable at the old rate. The investment function, given by the line $r_m I_m$ in Figure 1, will thus be negatively sloped.

This curve will shift to the right with an increase in income and employment. The reason can be seen from the properties of the aggregate production function, which specifies the quantity of output that can be produced in the economy for each level of labour and capital inputs. This function can be written

$$Y = F(N, K) \tag{3.1}$$

where Y is the level of output, N is the quantity of labour employed, and K is the quantity of capital employed. At any given level of technology, and given full employment of the existing capital stock, an increase in output can only occur if there is an increase in the level of employment of labour.

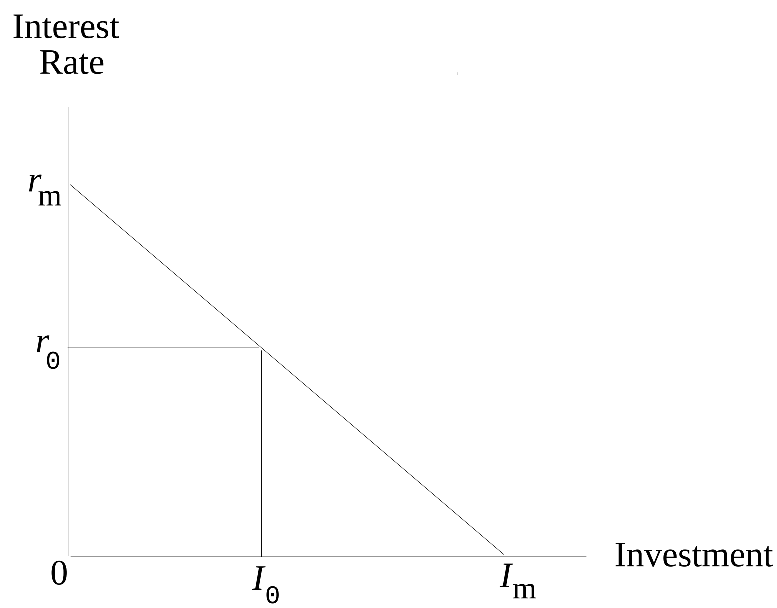

Figure 1: The Keynesian Invesment Function.

From the standard properties of production functions, we know that an increase in the labour input holding the capital input constant will lower the marginal physical product of labour and increase the marginal physical product of capital. Since a rise in the marginal product of capital in the economy raises the return to all investment projects, the rise in output and associated increase in employment shifts the curve $r_m I_m$ upward, increasing the number of investment projects that will be profitable at each rate of interest. Holding the market interest rate constant, therefore, an increase in output and employment will increase the equilibrium level of investment.

Keynes called the curve $r_m I_m$ the *marginal efficiency of investment*. Investment expands in the Keynesian model until the marginal efficiency of investment falls to equality with the market rate of interest.

3.1 Present Value vs. Internal Rate of Return

The above analysis contains an oversimplification that must now be corrected. Consider a firm with a two-period horizon, owned by an individual who also has a two-period horizon. The firm can produce Y_0^m units of output and income in year zero if all of its productive efforts are concentrated on that year. By channelling resources in to investment in new capital stock, however, and thereby reducing its production of final output in year zero, the firm can produce final output in year 1. Let the maximum output it can produce in year 1 if nothing is produced in year 0 be Y_1^m. The curve joining Y_0^m and Y_1^m in Figure 2 gives the firm's intertemporal production-income opportunities.

What output mix should the firm produce in the two years? To answer this question we might begin with the utility function of the owner of the firm. The indifference curve U_0, tangent to the production opportunity locus at point a and output mix (Y_0, Y_1), represents the maximum utility the firm's owners could obtain if they converted the outputs produced by the firm in each year into equivalent amounts of consumption in that year. In a world where borrowing and lending is possible, however, this does not represent an optimum. Suppose that consumers and firms can borrow and lend freely at the market interest rate r_0. Then the consumption possibilities of the owners of the firm at the output mix (Y_0, Y_1) are given a line through the point a with slope equal to $-(1 + r_0)$. On Figure 2, that line crosses the horizontal axis at P_V. Utility can be increased, however by shifting the output mix to point b and the consumption mix to point c along the line that crosses the horizontal axis at P_V'.

Because borrowing and lending is possible, there is a complete separation of the firm's intertemporal output production decision and the firm-owner's intertemporal consumption decision. In analyzing the optimal intertemporal output mix of the firm, we can thus ignore the utility function. The distance OP_V along the horizontal axis measures the present value of the point a output mix (Y_0, Y_1) in units of year 1 output. The present value of the point b output mix is equal to the distance OP_V'. These present values represent the amount someone would pay, in units of year 0 output, to purchase the firm. The object of the firm is to choose the intertemporal output point that maximizes its present value. This is the point where a line of slope $-(1 + r_0)$ is tangent to the production opportunity line. Optimization of production enables the firm's owner to choose her consumption mix along a more favourable intertemporal consumption opportunity line than the one passing through point a.

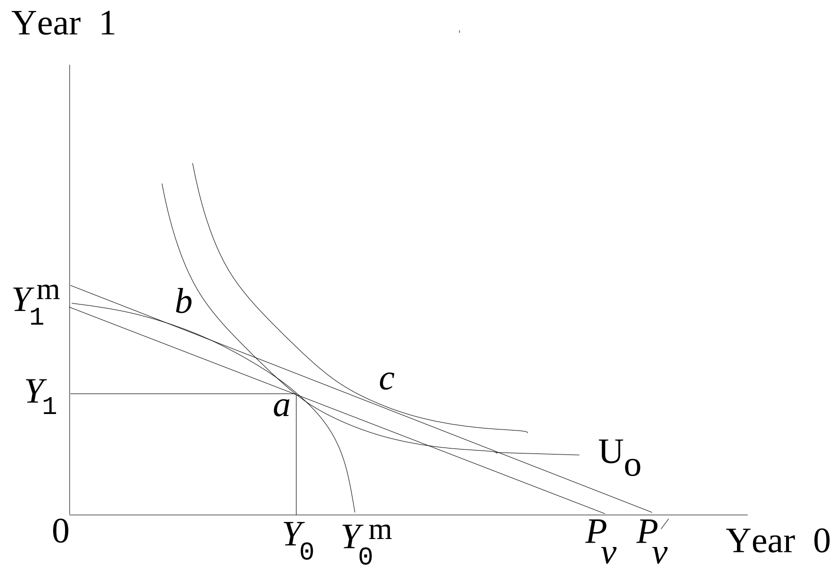

Figure 2: Intertemporal Production Opportunities of the Firm and Consumption Opportunities of its Owner.

The implication of this analysis is that firms will maximize their present value. They should therefore rank projects by their present value, which is not the same as ranking them according to their internal rates of return, as was done in the simple Keynesian presentation of the investment function above. Present value is calculated by discounting the excesses of returns over costs in the current and all future years by the market rate of interest according to the formula

$$P_v = R_0 - C_0 + \frac{R_1 - C_1}{1+r} + \frac{R_2 - C_2}{(1+r)^2} + \frac{R_3 - C_3}{(1+r)^3} + \ldots\ldots\ldots\ldots \qquad (3.2)$$

where R_t and C_t represent the returns and costs occurring in the t^{th} year. The internal rate of return on the project, represented in Figure 1 as the Keynesian marginal efficiency of capital, is calculated by finding the value of r in equation (3.2) that reduces P_v to zero.

It turns out that the present value and internal rate of return procedures do not yield the same ranking of projects. Consider two projects, each costing 1 unit in the year 0. Suppose that Project 1 yields no return in year 1 and a return of 4 in year 2, while Project 2 yields a return of 2 in year 1

Table 1. Internal Rate of Return and Present Value

Proj. No.	Cost Year 0	Return Year 1	Return Year 2	Internal R. of R.	Present Value r = 0	r = .5	r = 1
1	1	0	4	1	3	.78	0
2	1	2	1	1.414	2	.78	.25

Notes: The internal rates of return for the two projects are 1 and 1.414 as shown on the respective lines below

$-1 + 0/(1+1) + 4/(1+1)^2 = -1 + 0 + 4/4 = 0$

$-1 + 2/2.414 + 1/(2.414)^2 = -1 + .828500 + 1/5.827396 = -.1715 + .171603 \approx 0$

and the present values can be obtained by substituting the period returns and the market interest rate into the present value formula (3.2).

---

and a return of 1 in year 2. The calculations of internal rates of return and present values are shown in Table 1.[1] Project 2 ranks higher than Project 1 according to the internal rate of return criterion, while at market interest rates below 50% Project 1 ranks higher than Project 2 according to the present value criterion. The reason is that high interest rates discount very heavily the large returns from Project 1 in year 2, giving Project 2 the edge when the market interest rate is high. These same high returns in year 2 give Project 1 the edge when the market interest rate (and the resulting discount factor) is low. The internal rate of return criterion is inappropriate because it pays no attention to market interest rates in ranking projects. As Figure 2 indicates, it is present value that the firm should be trying to maximize.

Accordingly, the curve $r_m I_m$ in Figure 1 should be constructed by starting with a market interest rate at which no projects have positive present value, represented in the figure by the point r_m, and then progressively lowering the market interest rate and calculating the aggregate investment in all projects having positive present value at each market interest rate. The new curve will cross the horizontal axis at the point where all projects having positive present value at a zero market rate of interest have been undertaken. An increase in the level of output and employment in the economy will increase the present and future returns R_t, raising the present value of all projects at each market rate of interest. The curve $r_m I_m$ will shift

---

[1]This example is taken from William H. Branson and James M. Litvack, *Macroeconomics*, New York: Harper and Row, 1976, Chapter 2.

to the right. The use of the present value criterion in place of the internal rate of return in criterion to rank projects thus changes the construction of marginal efficiency of capital curve in Figure 1 but not the qualitative conclusions that emerge from the analysis.

3.2 The Capital Stock Market

Consider an economy having the aggregate production function given by equation (3.1) above. The wage of labour, measured in units of output, is the marginal physical product of labour, $\partial F/\partial N$, and the corresponding wage or rental rate of capital is $\partial F/\partial K$, the marginal physical product of capital. How much would someone be willing to pay to buy the rights to the future income from a unit of capital stock?

To simplify things, let us suppose that the flow of income from a unit of capital, represented by the current marginal physical product, is expected to continue at its current level indefinitely. Under these circumstances, the present value of this constant flow of future income is equal to $\partial F/\partial K$ divided by the real rate of interest. The price, measured in units of current output, of a unit of capital stock is thus

$$P_K = \frac{\partial F/\partial K}{r} \tag{3.3}$$

At this price, the marginal physical product of capital yields a return equal to the market interest rate.

The standard diminishing returns feature of the production function implies that an increase in the stock of capital holding the quantity of labour employed constant will cause the marginal physical product of capital to decline. Therefore, at any given interest rate an increase in the stock of capital will be associated with a decline in the price at which asset holders will be willing to hold that capital stock. This is indicated by the downward sloping demand for capital stock curve D_kD_k in Figure 3. A fall in the market real interest rate will, by lowering the denominator on the right hand side of equation (3.3), shift the D_kD_k curve upward and lead to an increase in the price of capital goods. A lower interest rate implies that the given real revenue or marginal physical product of a unit of capital will represent the going yield on a larger capital value. A rise in output and employment, holding the capital stock constant, will lead to an increase in the marginal physical product of capital, also shifting D_kD_k upward. The numerator of the right-hand side of equation (3.3) will increase and the present value and market price of a unit of capital stock will rise.

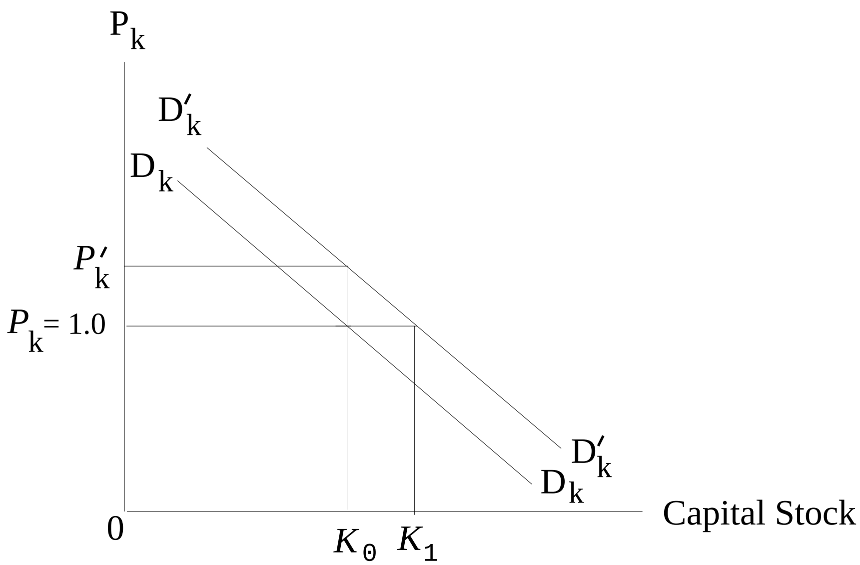

Figure 3: The Market for the Stock of Capital.

3.3 The Opportunity Cost of New Capital

If the market price or present value of capital happens to be above the cost of producing and installing an additional unit, there will be an incentive for firms to undertake investment. What, then, determines the cost of producing and installing new capital goods? To answer this question, consider Figure 4, which gives the economy's opportunities for producing new capital goods. If the entire output of the economy can be represented by a single aggregate good, used for both consumption and investment purposes, this transformation locus is a straight line with slope equal to -1 as shown by the solid line in the figure. One unit of new capital can be produced by sacrificing one unit of consumption because consumer and new capital goods are the same good. The level of income is equal to the sum of the quantities of consumer and new capital goods produced—it equals the distance along either axis between the origin and the intersection of the production opportunity line with that axis. The supply curve of new capital goods to the economy, as a function of the price of capital goods in output units, shown by the solid line in Figure 5, is perfectly elastic at a price of unity up to a level of investment equal to current output Y_0 and perfectly inelastic at that point.

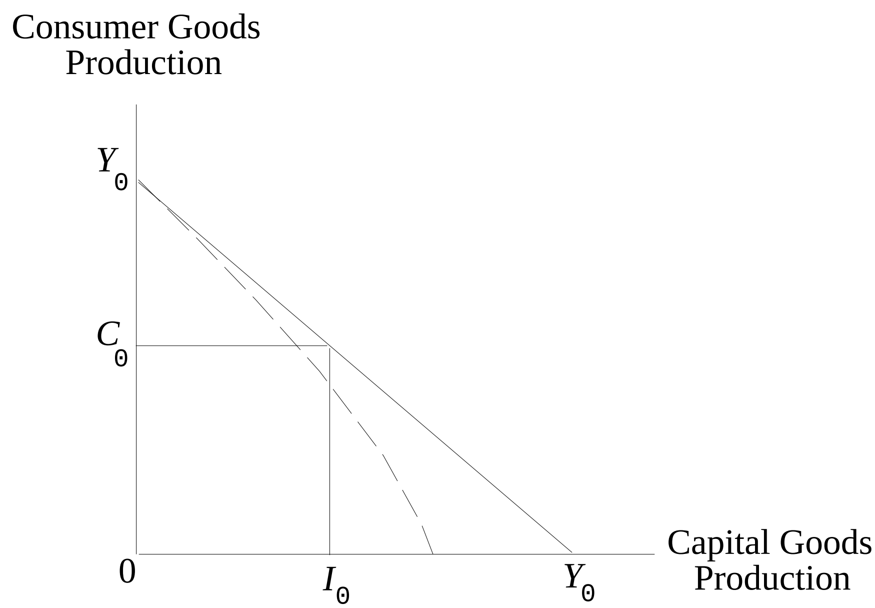

Figure 4: Capital/Consumption Goods Production Opportunities.

If the demand for capital stock is represented by the curve $D_k'D_k'$ in Figure 3 and the existing capital stock, K_0, is such that the present value of capital exceeds unity, there will be an incentive for new investment to take place. Indeed, if the distance between the current capital stock K_0 and the new equilibrium capital stock K_1 at $P_k = 1$ exceeds the current level of output Y_0, the entire output of the economy will be devoted to the production of new capital goods. Because the quantity of capital is many times larger than the flow of income from it, it will take only a small shift in the demand for capital to create a divergence between the actual and equilibrium stocks of capital greater than the level of output. This complete specialization in new capital goods production will continue until the capital stock has risen to the new equilibrium level K_1. Similarly, if through a rise in the interest rate or some other factor the demand for capital falls below D_kD_k, the price of capital will fall below unity and the entire output of the economy will be devoted to the production of consumer goods.

The investment function that results from this formulation is shown in Figure 6. It has a horizontal segment at the interest rate, represented by r_0, for which present value of a unit of capital is unity. When the interest rate

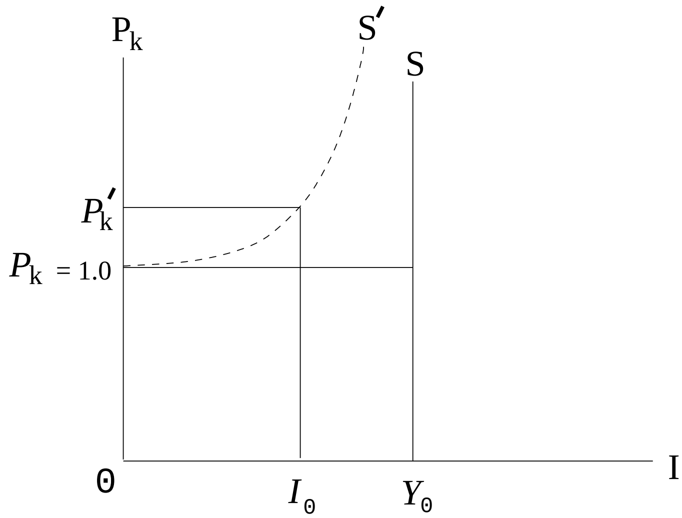

Figure 5: The Supply Conditions for New Capital.

rises above r_0 the present value of capital falls below unity and the level of investment goes to zero. When the interest rate falls below r_0, the present value of capital rises above unity and the level of investment increases to equal the level of output Y_0.

This implication that production will shift between complete specialization in new capital goods production and complete specialization in the production of consumer goods in response to small shifts in the demand for capital stock is grossly inconsistent with both common sense and the facts. Investment goods production is more variable than consumer goods production, as can be seen from Figure 7, which presents data for the United States in millions of 1982 U.S. dollars, and Figure 8, which presents the same series as deviations from their logarithmic trends. Gross private fixed investment and consumption of durable goods, which represents consumers' investment in automobiles, refrigerators, etc., are clearly more variable around their trends than consumption of non-durables and services. But it never happens that all the resources of the economy become devoted to the production of only investment goods or only consumer goods. The theory thus developed is thus inconsistent with the facts.

The notion that the economy's entire output could be devoted to invest-

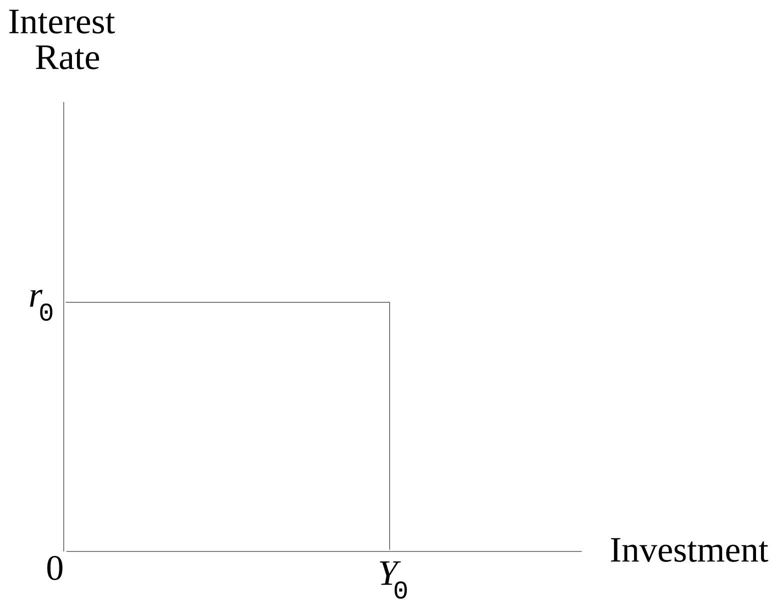

Figure 6: The Invesment Function when Capital Goods are Produced at Constant Cost in Terms of Consumer Goods.

ment goods production is also inconsistent with common sense.[2] Why would consumers willingly starve themselves? The interest rate typically falls in the above type of closed economy analysis because the government increases the money supply. The situation with respect to interest determination in the many-period case can be shown in Figure 9. Output in the current year, which also equals current year consumption in a closed economy when output is a perishable good, is shown on the horizontal axis and output in each future year is shown on the vertical axis. At a given stock of capital that never depreciates, output will equal Y in the current and all future years. The line rr, the slope of which equals the marginal product of capital, gives the tradeoff between current and perpetual future consumption. Its slope is also equal to the interest rate because a reduction of current-year consumption of 1 unit will increase perpetual future consumption by r units. The initial equilibrium is at point a on indifference curve U. The effect of a monetary expansion will be to increase the representative agent's holdings of monetary relative to non-monetary wealth at the existing interest rate,

[2] I would like to thank Nathan Nunn for comments that led to the analysis that follows.

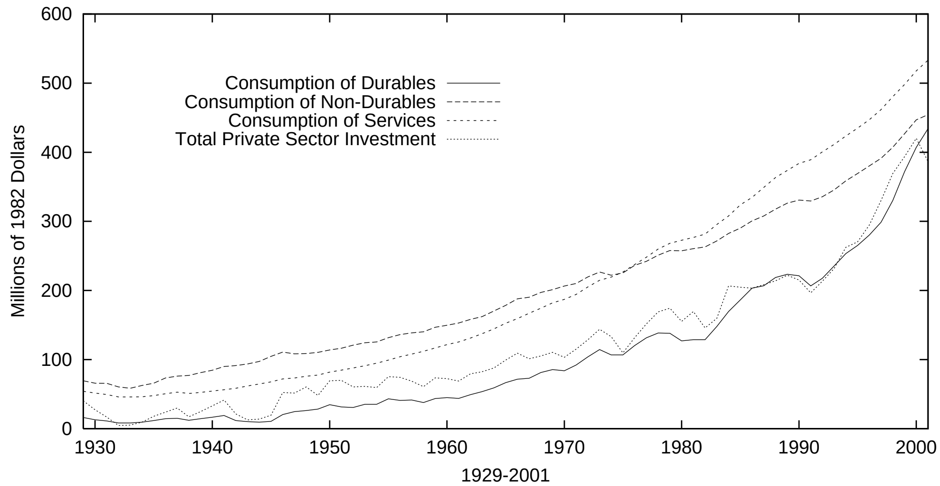

Figure 7: U. S. Personal Consumption of Durables, Non-Durables and Services, and Total Investment: Millions of 1982 Dollars. Source: *U. S. Historical Statistics* and *Citibank Data Base*.

causing him to rebalance his portfolio by trying to convert money into other assets. This will rotate his indifference curves on the plane shown in Figure 9 counterclockwise or shift them to the left. (Keep in mind that there is a third axis, representing monetary wealth, rising vertically from the plane at the origin.) The new indifference curve through point a will be flatter, signifying that the individual will hold the current mix of current and perpetual future consumption at a lower interest rate. But the interest rate is necessarily equal to the marginal product of capital because capital goods and consumer goods, being the same good, are produced at constant cost in terms of each other. So investment will become positive and the level of current consumption will decline, and perpetual future consumption will increase, to where the new indifference curve is tangent to the rr line. The monetary expansion will have no effect on the interest rate.

Monetary expansion can affect the interest rate only if it causes the indifference curve map to shift so much that the maximum utility occurs at the point where the rr line crosses the vertical axis on an indifference curve whose slope at that point is flatter than the rr line, a situation where the interest rate is below the marginal product of capital. In this case, consumers would be consuming nothing in the current year. But it is reasonable to expect that, since consumers will presumably not choose to starve themselves,

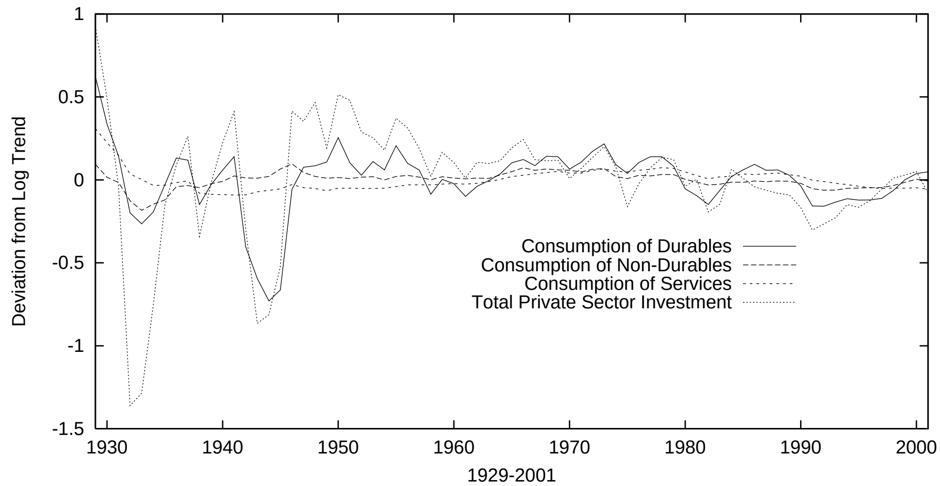

Figure 8: Deviation of U.S. Personal Consumption of Durables, Non-Durables and Services, and Total Investment from Log Trend. Source: See Table 7.

the indifference map on the plane in Figure 6 will never shift sufficiently to yield this result, no matter how large the stock of monetary wealth becomes. We must conclude that the observed interest rate will never change when capital goods are produced at constant cost in terms of consumer goods unless there is a change in the marginal product of capital.

3.4 Convex Production Opportunities

A more realistic portrayal of the investment process will arise if the production opportunity locus is convex, as in Figure 10. In that figure, new capital goods are produced at increasing cost in terms of consumer goods and the supply curve of investment takes a form given by the broken line in Figure 5. The rise in the price of capital goods from P_k to P_k' will result in an increase in the level of investment and a decline in consumer goods production but the equilibrium will be an interior rather than corner solution and all resources in the economy will never be devoted exclusively to either investment or consumption. Each year's investment will shift the vertical capital stock supply curve in Figure 3 to the right, lowering the price of capital and reducing the excess of the equilibrium over the actual capital stock. As the price of capital stock falls in response to successive periods'

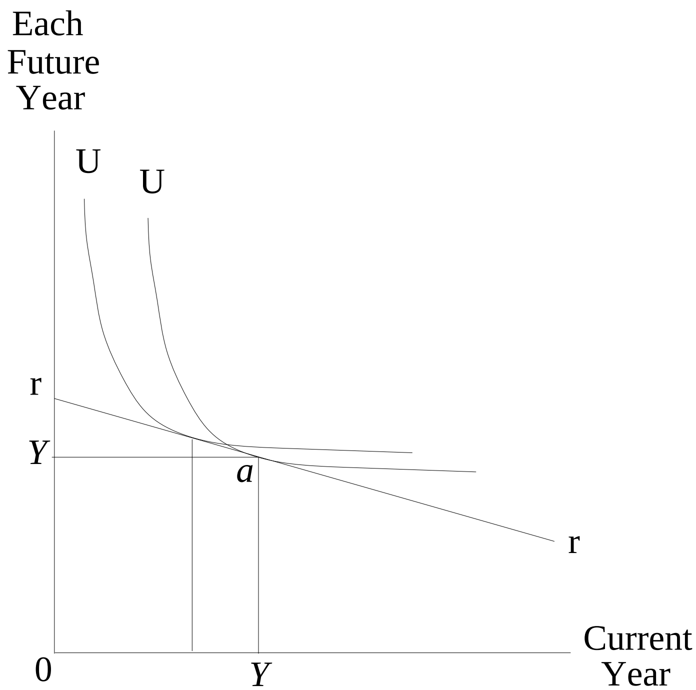

Figure 9: Interest Rate Determination When Capital and Consumer Goods are Produced at Constant Relative Cost.

investment, the level of investment declines, eventually approaching zero as the actual capital stock approaches the equilibrium one.[3]

The investment function that results from this formulation is negatively sloped and equivalent in form to the traditional one shown in Figure 1. A fall in the interest rate raises the present value of capital, stimulating increased production of new capital goods to the point where the marginal cost of production equals the present value. A rise in the level of income and employment raises the marginal product of capital, increasing its present value and stimulating an expansion of new capital goods production—the level of investment increases at each level of interest rates shifting the negatively

[3]The argument here assumes that there is no depreciation. Depreciation shifts the capital stock supply curve to the left at some rate through time, requiring a higher level of new investment than otherwise to maintain the same capital growth. In the presence of depreciation, a stationary capital stock will occur when the flow of investment equals the flow of depreciation.

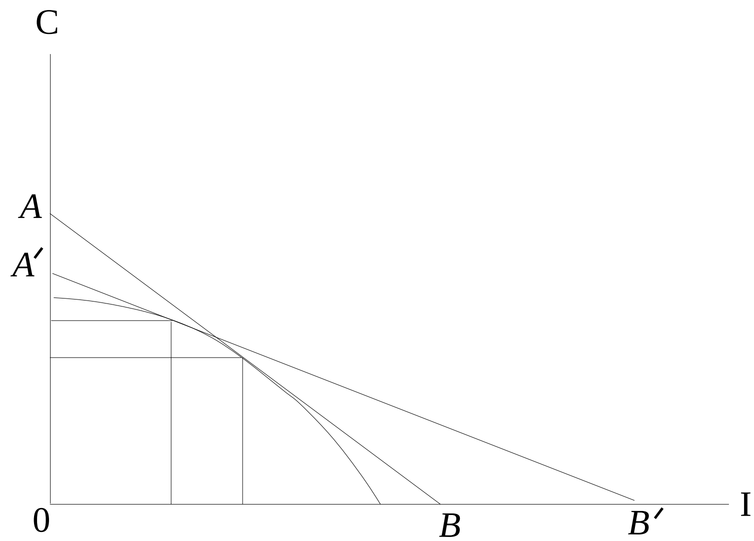

Figure 10: Convex Production Opportunities for New Capital Goods vs. Consumer Goods.

sloped curve to the right through time.

The assumption that new capital goods are produced in the economy at increasing cost in terms of consumer goods would seem to be a reasonable one in view of the fact that the different types of goods use different fixed factors of production—land in the case of food and some other consumer goods, and coal, iron, etc., in the case of capital goods. While it introduces an element of realism into the model, however, it also raises some additional problems.

Income cannot be uniquely measured when the production opportunity locus is convex as in Figure 10. At a relative of price of capital goods in terms of consumer goods, P_k, output equals the distance OA when measured in units of consumer goods and OB when measured in units of capital goods. If preferences should change causing the interest rate to rise, the relative price of capital goods to fall, and less new capital goods and more consumer goods to be produced, the level of output measured in units of consumer goods falls to OA' and the level of output measured in units of capital goods rises to OB'. Yet the economy's real opportunities for producing goods now and in the future have not changed. Though only the composition of output changes, our measures of the aggregate level of output change on account of changes in the relative price of capital goods in terms of consumer goods. This aggregation problem destroys the mathematical neatness of the analysis.

3.5 Adjustment Costs of New Investment

One way of dealing with this aggregation problem is to simply ignore it. Everyone knows that the entire output of the economy cannot be realistically aggregated into a single commodity but it is a convenient assumption. It is a minor further step to proceed as though our measure of output is unique. This creates a problem, however, when we try to proceed mathematically. Output changes not only as a result of changes in employment and the stock of capital—it changes every time the relative price of capital goods in terms of consumer goods changes.[4]

An alternative way to create an upward sloping supply curve of new investment is to assume that although new capital goods can be produced at constant cost in terms of consumer goods, there are additional costs of putting the new capital goods in place. Additions to the capital stock, so the story goes, involve dislocation as the new capital goods are adapted to the original capital stock and labour inputs. These adjustment costs are assumed to increase as the flow of new investment increases. The cost of capital in place, P_k, exceeds the cost of capital in production, call it C_k, by the amount $\lambda(I)\, C_k$, where $\lambda'(I) > 0$. Output is measured at the production level and is invariant to the mix of consumer and capital goods produced. Since the market price of new capital in place equals

$$P_k = C_k[1 + \lambda(I)], \tag{3.4}$$

we can rework equation (3.3) to express the interest rate as

$$r = \frac{\partial F/\partial K}{1 + \lambda(I)} \tag{3.5}$$

where C_k is normalized at unity. Since $\lambda'(I)$ is positive, an increase in investment increases the denominator in equation (3.5), reducing the interest

[4]Another problem is that a movement along the production opportunity curve will lead to a change in the marginal productivity of capital in the economy if the capital/labour ratio in the production of new capital goods differs from that in the production of consumer goods. For a discussion of these issues, see J. E. Floyd and J. Allan Hynes, "Capital Immobility, Adjustment Costs, and the Theoretical Foundations of Income-Expenditure Models," *Journal of Political Economy*, December 1978.

rate. A negative relationship between the interest rate and level of investment of the form portrayed in Figure 1 is established. An increase in output raises the marginal physical product of capital in the numerator, shifting the downward sloping investment function to the right in that figure, just as in the case where the investment function was derived on the basis of convex production opportunities.

The adjustment cost approach rationalizes the standard investment function while at the same time maintaining the fiction that output can be uniquely defined as the sum of consumer and new capital goods production. This is a fiction because though goods output can be uniquely measured, it no longer measures the true output of the economy. The adjustment costs of putting new investment in place involve a reduction of the output that can be produced with the existing labour and capital resources. Our measure of output conveniently ignores these resources used up in adapting the new investment goods to the old labour and capital inputs. If we count these resources, the level of output when properly measured declines as a consequence of an expansion of the level of investment. The 'true' production opportunities are represented by the dotted line in Figure 4.

3.6 The User Cost of Capital, Liquidity Constraints, and Tobin's q

The equilibrium level of investment can also be analyzed in terms of the return of and cost of capital to the firm. The return to adding another unit of capital is the marginal product of capital $\partial F/\partial K$. The cost of capital, called the *user cost*, is composed of the alternative opportunity cost of the funds that must be used to buy the capital and put it in place plus the funds that must be set aside to replace the capital that has depreciated. The former is represented by the interest rate times the cost of purchasing new capital goods and get them working, rP_k, while the latter is represented by $\delta\, C_k$ where δ is the depreciation rate per unit of capital. The cost of depreciated capital is simply its cost of production because the adjustment costs of putting new capital in place have already been borne for capital being replaced. In a world where the present value of capital is expected to increase through time, the rate of increase of the price of capital yields a capital gain to the firm. This capital gain, taken as a percentage of the value of capital, $\partial P_k/(P_k \partial t)$, can be treated as a reduction in the firm's cost of capital. The user cost of capital can thus be written as

$$rP_k + \delta C_k - \left[\frac{1}{P_k}\frac{\partial P_k}{\partial t}\right] P_k.$$

As shown in the previous section, the market value of capital, P_k, will be related to the cost of producing it according to

$$P_k = C_k[1 + \lambda(I)].$$

The firm will operate where the return to adding a unit of capital stock equals costs of putting it in place minus the expected capital gain. Normalizing the constant cost of producing new capital goods in units of consumer goods, C_k, at unity, we can express the firm's equilibrium condition as

$$\begin{aligned} \frac{\partial F}{\partial K} &= r\,[1 + \lambda(I)]\,C_k + \delta\,C_k - \left[\frac{1}{P_k}\frac{\partial P_k}{\partial t}\right] P_k \\ &= r\,[1 + \lambda(I)] + \delta - \frac{\partial P_k}{\partial t} \end{aligned} \tag{3.6}$$

which reduces to

$$r = \frac{\partial F/\partial K - \delta + \partial P_k/\partial t}{1 + \lambda(I)}. \tag{3.7}$$

This equation equals (3.5) when the depreciation rate and the expected capital gain from holding capital are zero.

The ratio of the market value of capital to its cost of production is known as Tobin's q.

$$q = \frac{P_k}{C_k} = \frac{\partial F/\partial K - \delta + \partial P_k/\partial t}{r\,C_k} = 1 + \lambda(I) \tag{3.8}$$

Since the adjustment costs of putting new capital in place are positively related to the level of investment, q varies directly with the investment level.

It is frequently argued that funds generated within the firm as a result of profits are a cheaper source of investment finance than funds obtained by borrowing from financial institutions or issuing bonds or equity shares. This implies that the interest rate the firm must pay to borrow funds exceeds the interest rate at which it can lend them out, which would mean that the returns to the firm from investing its profits internally are greater than the returns from investing elsewhere in the economy. The question immediately arises as to why market investors would turn down excess profit opportunities from investing in the firm rather than elsewhere at market rates. Does the market would regard the firm's projects as more risky than diversified investments elsewhere? If so, why would the firm invest in itself at high risk when other investors are not willing to do so? It must be that the firm has information about itself that other investors do not have.

To the extent that the borrowing rate of the firm for the firm is above the internal cost of capital for projects that are no more risky than other projects in the economy, the cost of capital of the firm must be calculated using an interest rate internal to the firm in place of the market interest rate. This implies that the present value of the firm's capital to its owners is higher than the present value of that capital to investors outside the firm. The firm will add to its capital stock to the point where the internal present value is equated to the sum of the cost of producing new capital goods and the adjustment costs of installing them.

3.7 The Dynamics of Investment: Persistence and the Accelerator

The above analysis outlines various derivations of the investment function

$$I = I(r, Y) \tag{3.9}$$

which is the main vehicle for incorporating investment into standard traditional aggregative macroeconomic models. A number of embellishments are often used within this framework to interpret actual movements of investment.

First, it must be noted that the building of an investment project is not simply a one-year deal. The firm must plan its long-term investment strategy because many types of capital do not depreciate rapidly—once in place, they remain for a long time. Moreover, with the invention of new technology, capital equipment becomes obsolete—that is, its returns decline relative to the returns on investment in newly designed equipment that does the same job, and its present value falls below replacement cost. Once an investment decision has been made, construction may take several years and it may be extremely costly for the firm to reverse its investment decision and abandon the project. Investment taking place in any one year, therefore, will represent the result of investment decisions that were made in a number of preceding years.

There are two general ways to incorporate this 'time to build' feature in the investment function. One is to note that the interest rate relevant for this year's investment is not just the current interest rate, but the interest rates that ruled in previous years when the decisions were made to construct the projects that are now under construction. This would require the addition of one or more lagged interest rates, r_{t-i}, as arguments in the investment function in equation (3.9). A second way to incorporate the lagged effects on current investment of past investment decisions is to note that last year's investment, I_{t-1}, reflects all decisions made last year and previously about the level of investment. The effects on current investment of all those decisions can thus be taken into account by adding I_{t-1} as an argument in the investment function in equation (3.9) instead of including several years' lagged interest rates. Whichever approach is used, the effect will be to introduce persistence into the movements in the endogenous variables through time generated by the macroeconomic model in which the investment function is included. Since a fall in the interest rate in a particular year leads to an increase in investment not only in that year but in several subsequent years, the higher level of income resulting from the fall in the interest rate will persist some years into the future. These conditions will thus lead to persistence in movements of output and income through time.

'Time to build' is not a feature of all investment decisions. Houses and buildings are frequently built in less than a year and inventories can be accumulated and drawn down within a much shorter period. It is nevertheless an important characteristic of a substantial part of the ongoing investment flow in most economies.

A second frequently made embellishment of the investment function is the 'accelerator'. It has been noted that an increase in output, by increasing the marginal product of capital, shifts the stock demand curve for capital outward to the right, increasing the desired capital stock relative to the actual stock. In the simplest case, where the shift in demand is small enough, investment goods are produced at constant cost in terms of consumer goods, and adjustment costs are zero, this entire increase in desired capital stock will be taken care of by current investment. The level of investment can thus be expressed in this simple case as

$$I = K_t - K_{t-1} = K_t^* - K_{t-1} = \alpha\, Y_t - \alpha\, Y_{t-1} = \alpha\, \frac{dY}{dt} \qquad (3.10)$$

where K_t^* is the desired capital stock in year t, equal to the actual stock on account of current investment, and α gives the effect of aggregate output on the desired capital stock via its effect on the marginal product of capital.

The implication of equation (3.10) is that the level of investment will be determined by the rate of growth of output. An exogenous aggregate demand shock will have an accelerating effect on current output as a result of the fact that the level of investment is directly related to the rate of growth of output. This formulation can be used to generate a variety of dynamic models of the investment process and output growth. Normally, of course, it is recognized that interest rates and the level of output will also have an effect on investment independently of the rate of output growth. So the acceleration principle is introduced by including the rate of growth of output as an argument, along with r and Y, in the investment function in (3.9).

The acceleration principle must be used very cautiously. It is reasonable to believe that the firm will have long-term investment plans that reflect its beliefs about the rates of growth of output that will occur in future years as well as the current year. A change in output growth in the current year will not necessarily modify the firm's beliefs about the future growth of output at all—it will depend on whether the change in output growth is a permanent or transitory phenomenon. Moreover, if the economy is a full employment, a change in output will be accompanied either by an increase in technology or an increase in the fully employed labour force or both. Since the capital/labour ratio may not be changing, it is by no means assured that the increase in output will be associated with an increase in the present value of a unit of capital above the level that firm had previously expected on the basis of its investment plans and its knowledge of the future. If the present value of capital does not rise above the level anticipated, the desired capital stock will remain the same as anticipated, and the previously planned level of investment for the period will turn out to be the optimal level of investment. Clearly, it is not reasonable to assume that, at all times and places, an increase in the rate of output growth will lead to an increase in the level of investment at current output and interest rate levels. In modelling the dynamics of the investment process, we should probably give no more attention to the rate of output growth than to a number of other possible variables that could characterize firms' expectations as to the future returns to capital.

In addition, it should be noted that when capital is broadly defined to include human skills and technological and more basic knowledge, as consistent with the growth rate being endogenous—that is, a function of the investment in physical capital, human capital, technological innovations and basic knowledge—output can grow only as a result of the growth of broadly defined capital. In this framework, it makes little sense to think of an exogenous increase in the growth rate of output as an increase in the level of the return from capital.

Quite apart from the difficulties noted above, the rate of growth of output may have an important relationship to inventory investment. Although modelling of the dynamics of inventory investment is a complex undertaking, it turns out that an increased rate of growth of aggregate demand and output will result in an increase in the desired the level of inventories. There may thus be a direct relationship between output growth and the level of investment in inventory accumulation. This is not, however, the only factor determining inventory investment. Interest rates are important, as are firms' forecasts about the level and variance of future demand for their products.

3.8 Empirical Evidence on the Determinants of Investment

A substantial amount of empirical work has been undertaken over the years to verify the general characteristics of the investment function outlined in the analysis above. The early work is summarized by D. W. Jorgenson, who was also a major contributor.[5] Charles Bishoff found that the long-run elasticity of investment demand in the United States with respect to changes in output is unity and that the long-run elasticity with respect to changes in the interest rate is about -.5. He also found, as did earlier work, that changes in interest rates affect investment with a lag.[6]

There have also been efforts to measure Tobin's q and verify the relationship between it and the level of investment.[7] Since q is the ratio of the present value of capital to its production cost, one way of measuring it is to compare an index of the value of firms' shares on the New York stock exchange with an index of the prices of new capital goods.

A final related issue concerns the effect of the relation between income and investment on the slope of the IS curve. Consider the equation represented by that curve

$$Y = C(r, Y) + I(r, Y) + G \tag{3.11}$$

[5] See D. W. Jorgenson, "Econometric Studies of Investment Behaviour: A Survey," *Journal of Economic Literature*, December 1971, and comments by R. Eisner and L. R. Klein, *Journal of Economic Literature*, March 1974.

[6] C. W. Bishoff, "Business Investment in the 1970's: A Comparison of Models", *Brookings Papers on Economic Activity*, Vol. 1, 1979.

[7] See James Tobin and William Brainard, "Asset Markets and the Cost of Capital", in Bela Balassa and Richard Nelson, eds. *Economic Progress, Private Values and Public Policy: Essays in Honour of William Fellner*, North Holland, 1977.

and differentiate it totally to yield

$$dY = \frac{\partial C}{\partial r}dr + \frac{\partial C}{\partial Y}dY + \frac{\partial I}{\partial r}dr + \frac{\partial I}{\partial Y}dY + dG, \tag{3.12}$$

which can be consolidated into the form

$$Y = \frac{\partial C/\partial r + \partial I/\partial r}{1 - \partial C/\partial Y - \partial I/\partial Y}\, dr + \frac{1}{1 - \partial C/\partial Y - \partial I/\partial Y}\, dG \tag{3.13}$$

The IS curve will be negatively sloped if the coefficient of dr in the first term to the right of the equality is negative. For this to be true, the sum of the partial derivatives of consumption and investment with respect to the rate of interest in the numerator must be negative and the term in the denominator must be positive. Since it cannot be established that interest rate changes have a significant positive effect on consumption and since the interest elasticity of investment in the long-run is negative, it seems reasonable to suppose that the sum of the numerator terms is negative. For the denominator to be positive, the sum of the marginal propensities to consume and invest must be less than unity. The marginal propensity to consume in the U. S. appears to be in the order of .60 to .65. Since the income elasticity of investment is about unity, it follows (from the definition of income elasticity) that the marginal propensity to invest must be about the same as the ratio of investment to GNP, which is about 0.15. The denominators in the expressions to the right of the equality in equation (3.13) would appear to be clearly positive. This establishes that the assumption of a negatively sloped IS curve in aggregative macroeconomic models is a reasonable one.

3.9 Exercises

1. What is meant by the firm's intertemporal income possibility curve? How is it derived? What determines the firm's equilibrium position on that curve?

2. Consider the owner of a firm who has to make a production decision regarding how to distribute the firm's potential two-period earnings from production across the two years. How does the owner's utility function enter into this decision? What role does the interest rate play in this process?

3. What is the present value criterion? How does it relate to the firm's intertemporal production-income possibility curve?

4. What is meant by the internal rate of return criterion? How does it relate to the firm's intertemporal production-income possibility curve? Does it lead to the same ranking of projects as the present value criterion?

5. What is meant by the user cost of capital? What is its relationship to the rental rate on capital equipment? What is its relationship to the rate of interest?

6. What is meant by the accelerator principle? What role does it have in generating business cycles?

7. What is the relationship of the investment function to the marginal productivity of capital? To the IS curve?

8. True or False: Explain your answer briefly.

a) In order to manage a firm effectively, a manager must know the utility function of the owner(s) of the firm.

b) It makes no difference whether the firm plows its profits back into investment or borrows the funds to finance its capital investment because the opportunity cost of investment is the rate of interest in both cases.

c) The investment function is negatively sloped because the marginal product of capital declines as the stock of capital increases.