NOTES ON LABOR ECONOMICS

Economics 150

January 2000

William A. Sundstrom

Department of Economics

Santa Clara University

Contents

Chapter 1: Introduction

This is a course in the economics of labor markets. Labor economics encompasses many of the most important issues in economics. Most people earn most of their income by selling their labor time. So labor economics deals with the major source of personal income, what determines it, and why it may differ for different individuals. It also deals with the allocation of the most important (in value terms) input into the production process.

Some examples of important issues we will consider in the course are:

- Minimum wage/living wage: The current U.S. minimum wage is $5.15 per hour. In California, $5.75, and under San Jose's "living wage", workers on city contracts are supposed to be paid $9.50. Does the minimum wage cost jobs? Does it reduce poverty?
- Does college "pay"? Given the direct and opportunity costs of attending college, what is the rate of return on your investment?
- Effects of immigration and globalization: Some have argued that recent waves of low-skilled immigrants and competition from cheap imports have reduced the pay of low-skilled U.S. citizens. Is this true? How big an effect?

We are studying the economics of the labor market, but what exactly is "the labor market"? We can define it simply as all the buyers and sellers of labor services, and the institutions the facilitate that buying and selling. But what are labor services? Labor services are the direct input of human muscle and brainpower into production. Obviously this is very broadly defined, and

includes many occupations and tasks.

Labor services are distinct from the other major types of inputs, such as raw materials or capital services. But we'll see that this distinction is not so clear, for we will make use of the idea that people invest in their own skills, which we will think of as *human capital*.

In fact the labor market consists of many markets. Labor markets differ in terms of location, occupation, and skill. A labor market tends to be more like a single market to the extent that there is a high degree of mobility within it. In this sense McDonald's and Burger King servers are probably in the same labor market, but fast-food workers and heart surgeons are not really in the same labor market at all.

Keep in mind that not all work is bought and sold in labor markets. Much housework and child care labor, for example, is performed by family members without pay. Such work is not paid for in a market, but it may have important implications for people's behavior in markets for paid labor, because it constitutes an alternative use of a person's time.

Quantity and price in the labor market

As in any market, we are interested in quantity and price. The quantity is the amount of labor input, typically measured in number of workers and/or amount of labor time (say worker hours). The price is typically measured by the amount workers are paid per unit of labor time: the wage, or earnings. But this is also subtle, since workers care about more than just the pay of their job. There are non-wage benefits, and working conditions.

Let's start with some simple definitions relating to the quantity and price of labor. Quantity first: How many workers are there in the market? This is the question of the *labor market status* of the population. How much labor does each worker supply? This is the question of *hours of labor*.

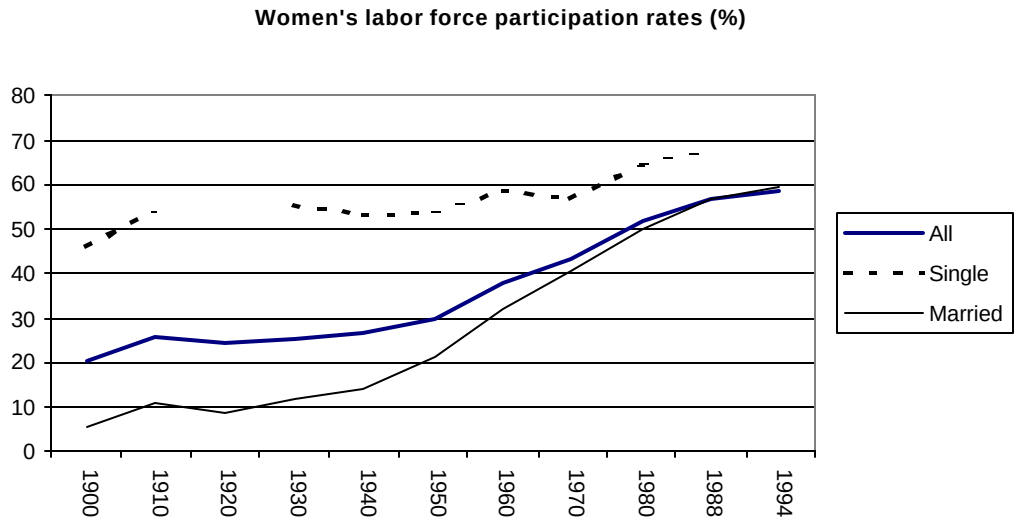
Measures of labor market status and hours

- *Labor force*: employed, actively seeking work, or expecting recall from layoff. The U.S. labor force is currently about 140 million.
- Everyone else is not in the labor force (*NILF*).
- *Labor force participation (LFP) rate* = labor force ÷ population. In the U.S. this is around 2/3 (67%) for adults. We can measure this for various demographic groups.

- *Employed*: at work for pay or profit. Note: Neither labor force nor employment includes people whose only work is household work for the family. This is a major shortcoming of our economic concepts, and it is paralleled by the problem that we do not measure household production as part of GNP.
- *Unemployed*: not currently working, but in the labor force (actively seeking work, or expecting recall from layoff)
- *Unemployment rate* = number unemployed ÷ labor force = just over 4% currently
- *Employment-population ratio* = number employed ÷ population
- *Hours of labor*: hours spent at labor activity (usually paid hours) per week. There are many difficulties in measuring this accurately.

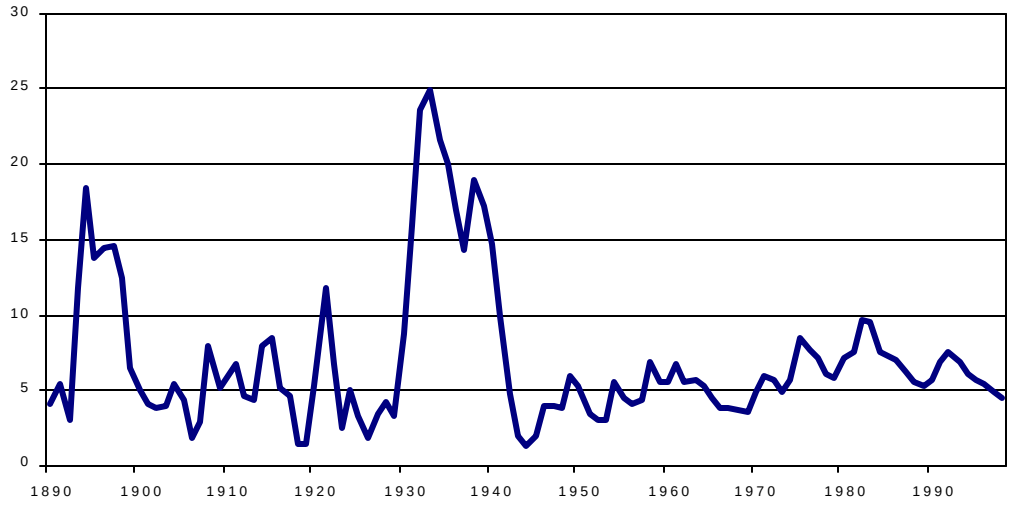
Some significant changes in labor market status over time

Perhaps the most dramatic changes in labor market status have been changing LFP rates by gender and age. Participation rates of women (especially married women) have risen dramatically since 1950 (see figure). This does *not* necessarily mean more married women are working, but rather that more married women are working *outside the home*. On the other hand, the participation of younger and older persons has fallen in recent decades, because of increases school attendance and earlier retirement.



Another interesting feature of labor market status over time is the cyclicality of unemployment rates (see figure). At present unemployment is at its lowest in decades, and one doesn't hear so much about unemployment lately, except in Europe, where it is much higher.

Civilian unemployment rate (%), 1890-1998



Measures of pay

- *Wage rate*: pay per unit of time worked (e.g. $12/hr.). The wage rate can be defined even for those not paid on an hourly basis by taking their salary and dividing it by their work hours.
- *Earnings* = wage rate × units of time worked. For instance, if the wage rate is $12/hr. and the worker works a 40 hour week, weekly earnings are $480. If the worker then worked 50 weeks in the year, annual earnings would be $24,000.
- *Total compensation* = earnings + other payments, such as fringe benefits
- *Income* = total compensation + "unearned" (non-labor) income. The latter could include interest earned on investments, gifts, inheritances, government transfers, etc.
- Nominal vs. real pay: In this course we will typically be interested in real wages and earnings, after taking out the effects of price inflation.

Recent wage trends

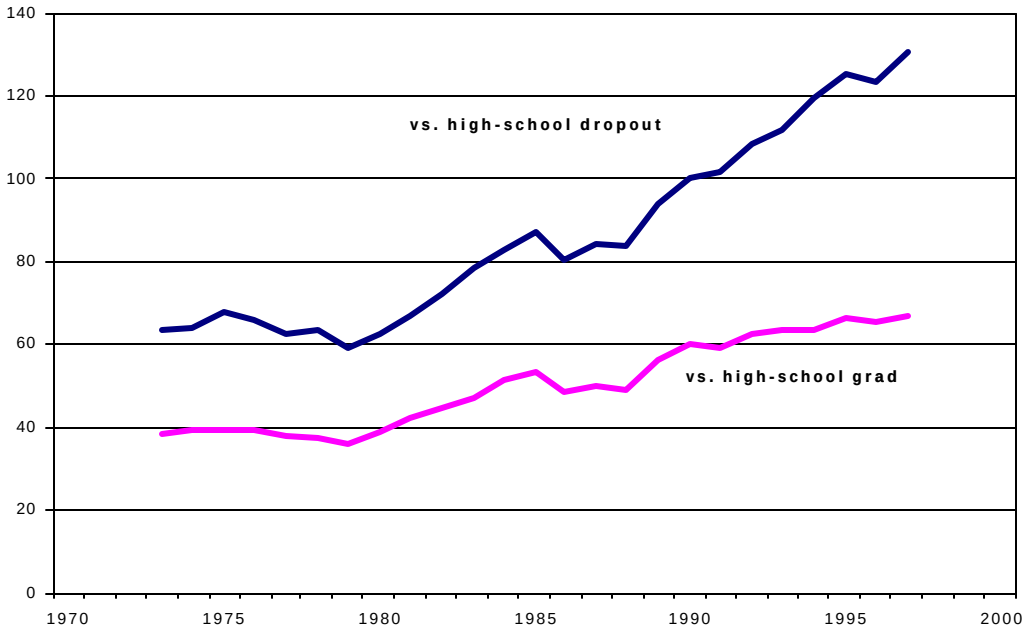
We will examine recent trends in the level and inequality of wages in much greater detail later in the course. Currently the median hourly wage is around $11.00. For a full-time year-round worker, this would yield an annual income of about $22,000. By conventional measures the last 20-30 years saw a significant stagnation of real average hourly earnings. But note that how one controls for inflation matters a lot here. Many economists consider the CPI deflator to overstate inflation; using a more accurate estimate of inflation shows real wage growth slowing but continuing upward after the 1970s.

The following table shows average wages over the last 20 years. The last row uses the modification of the CPI price index suggested by the Boskin Commission as a way to reduce the bias in the measure of inflation. See if you can figure out how each number in the third through fifth rows was calculated. Also, why is the average (mean) wage in 1997 (about $12) higher than the median (about $11)?

Average hourly earnings (U.S. private-sector nonsupervisory workers)

	1977	1987	1997
Average hourly earnings	$5.25	$8.98	$12.26
Consumer price index (CPI), using 1982-84 as base year	63.2	113.6	160.5
Average hourly earnings, 1982-84 dollars (using CPI)	$8.31	$7.90	$7.64
Average hourly earnings, 1997 dollars (using CPI)	$13.33	$12.69	$12.26
Average hourly earnings, 1997 dollars (using CPI minus 1 percent per year)	$11.00	$11.51	$12.26

Another significant trend during the same period is increasing *wage inequality*: a growing gap between "haves" and "have-nots." There are various ways to measure this, but one is to examine trends in wages by skill level, such as the education of the worker. The following figure shows the "wage premium" for male workers with four years of college, relative to high-school grads and men with less than a high-school education.



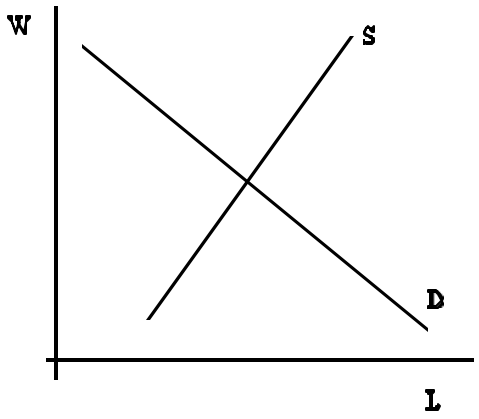
Wage premium of male worker with college degree (percent)

While overall inequality has increased, the historical gap between men and women's wages has narrowed, but is still large. In 1997, adult women who worked full-time earned about 67 percent of what adult men did, up from about 50 percent in 1970. Later in the course we will examine some potential sources of this gender differential and reasons for its narrowing.

Chapter 2: Supply and demand in the labor market

In much of this course we will be analyzing labor supply and labor demand in some depth. To place the discussion in perspective, here's a quick refresher on supply and demand analysis, in this case applied to labor markets.

The basic analysis is similar to what you've had in many economics courses. But remember that here we are looking at an *input* market. So the demand side is the demand by producers (business firms and other employers). The supply side is supply by all the individual workers. Keep in mind that supply and demand analysis is only appropriate for competitive markets. It does not work well to describe cases where employers have monopsony power or workers have monopoly power.



Your old friends, supply and demand

Demand for labor

The *labor demand* curve shows the quantity of labor demanded by employers at each given wage rate, *holding all other variables constant*. Demand is by firms (employers). We can speak of one firm's demand for labor, and then market demand by all the firms together.

Why does a firm's labor demand curve slope down? That is, why does an increase in the wage reduce the desired employment level by the firm? We will see that there are two important ways in which a change in the wage affects the quantity of labor demanded:

- *Scale or output effect*: As wage increases, costs increase, and the firm will typically choose to produce less. Therefore it needs less labor.
- *Substitution effect*: As wage increases, firm may try to substitute other inputs for labor (e.g. machinery or other K goods). Again this reduces the quantity of L demanded.

What could cause *shifts* in a firm's labor demand?

- *Change in product price*: If product demand increases, the product price will rise, and the firm wants to produce more at any given wage, so labor demand increases (shifts right). In this sense economists refer to labor demand as a derived demand, for the demand for any input is derived from the demand for the product of that input. Changes in demand for the product can arise for a variety of reasons, including changes in consumer incomes or tastes, changes in the prices of other goods consumers purchase, business cycle conditions.
- *Changes in other input prices*: These too have scale and substitutions effects, but in this case they can be offsetting. For example, suppose computers get cheaper. How will this affect the demand for clerical workers? On the one hand, the substitution effect may well be that the firm substitutes computers for clerks. But on the other hand, cheaper computers lowers costs overall, allowing the firm to produce more and therefore demand more of all inputs. Therefore, we cannot predict which way the demand for clerks will go overall.

In analyzing the determination of price and quantity in the entire market we would want to add up the demand curves across all the firms hiring in this labor market to obtain the market demand curve. We can expect that the market demand will be downward-sloping for the same reasons that individual firms' demand curves are.

Another important issue in labor demand is the time horizon: the distinction between short-run and long-run. Firms can make more adjustments in the long run, and thus we will see that there is some reason to believe that labor demand is more elastic in the long run.

Supply of labor

The *labor supply* curve is the quantity of labor supplied by workers at each given wage rate, *holding other variables constant*.

We can start by thinking about the supply of labor by an individual. It is often convenient to think of the individual labor supply decision in two parts, although they are related:

- *Participation decision*: whether or not to work any hours. We will see that participation generally depends on how much the worker could earn. A worker's *reservation wage* is defined as the wage rate below which an individual chooses not to work at all.
- *Hours decision*: how many hours to work, if working. When the wage increases, it has an ambiguous effect on each worker's hours. The higher pay increases the attractiveness of working for pay relative to other uses of time, but it also raises income and thus may make taking it easy more affordable. We'll come back to these effects later in the course.

Putting these together we can draw an individual supply curve that starts at a positive wage with hours increasing in the wage at first, but perhaps bending back at a high enough wage.

Market labor supply is the sum of the hours supplied by all the individual workers at each wage rate. How does a wage change affect total hours of labor supplied to the market? First, a rising wage brings more workers into the market (new entrants or job switchers), as the wage is above the reservation wage for a larger number of workers. But second, a rising wage has ambiguous effect on each worker's hours. Thus overall the effect is ambiguous. That is, the labor supply curve does not necessarily slope up.

What could causes of *shifts* in labor supply?

- *Changes in wages for other (competing) jobs*: Increase in other wages should decrease supply to a labor market as workers see alternative jobs as more attractive.
- *Changes in non-labor income*: This decreases supply, as workers can afford to work fewer hours, and some may even leave the labor force.

- *Changes in tastes or social structure*: for instance, values regarding the appropriateness of married women working outside the home.
- *Demographic shifts*: Because labor supply varies with age and other characteristics, market supply can shift as the composition of the population changes. For example, a population with a growing percentage of very old people is likely to see its labor supply decrease or grow less rapidly.

Equilibrium of supply and demand in the labor market

Of course the equilibrium in a competitive market is where the curves cross. We an use the diagram to analyze many changes in labor markets which shift the curves around. For example, in class we will consider various possible effects of "globalization."

We can also use the analysis to look at the effects of policy interventions in the labor market, such as the minimum wage. The current federal minimum wage is part of the Fair Labor Standards Act, which dates to 1938 and sets both minimum wages and maximum working hours (before overtime must be paid). The minimum wage is *binding* if it is set at a level above where the free-market equilibrium would be.

Other examples of government regulations and policies that affect labor markets include taxes on labor income (income and payroll taxes), transfer payments (such as social security and welfare), health and safety (OSHA), child labor law, etc. We will examine a number of these over the course of the quarter.

Imperfect competition in the labor market

There are many possible reasons that the competitive model of supply and demand is not appropriate for analyzing real-world labor markets. Some sources of imperfect competition include:

- *Unions*: Unions can have various effects, but they may be able to impose a wage above the competitive equilibrium, thus acting as a *cartel* or *monopoly*
- *Monopsony*: single employer. A famous example is the "company towns" Professional sports before free agency.
- *Information problems*: Supply and demand models break down if there are significant asymmetries in information. For example, it is often hard for employers to observe the

abilities or work effort of workers. Then workers may get stuck at one employer (can't easily prove their value to another); employers may pay efficiency (above-equilibrium) wages as an incentive to keep good workers or encourage greater effort.

- *Wage rigidity:* Wages tend to be downward sticky– it's hard for employers to cut them, even when demand falls.

Chapter 3: Labor supply and the theory of time allocation

The supply of labor to the market reflects the decisions of many individuals. Each person has the following basic decisions to make:

- *The labor-force participation decision:* To work in the market or not?
- *The hours decision:* If working, then how many hours per day?

Both these decisions can be thought of as decisions regarding the allocation of time. In labor economics, we tend to think of this as involving the tradeoff between two goods: "leisure" time and consumption. The tradeoff exists because the more leisure time you take, the less income you earn that can be spent on consumption goods. Of course, what we call "leisure" could actually include a variety of activities that are alternatives to paid employment, including unpaid labor in the household.

In this chapter we develop a basic model of time allocation, with the goal of understanding the effect of changes in wages and incomes on an individual's decisions. Then in the next chapter we explore some extensions and applications of the theory to such areas as the effects of alternative tax and transfer policies on work behavior.

Individual labor supply: basic elements of the theory

Our theory of labor supply starts with a single individual deciding whether to work for pay and if so how many hours. There are three basic elements of the theory:

- *Budget constraint:* What options are available to the individual? That is, what is the tradeoff between more leisure and more income? This depends mostly on the wage

rate per hour of work and any non-labor income the person has.

- *Preferences (indifference curves):* What are the individual's preferences or tastes regarding income versus leisure?
- *Optimal choice:* Given the available options, what choice of time allocation is most preferred by the individual?

The budget constraint

Let's start by looking at the budget constraint, which tells us the options available to the individual given the prices of goods and the wage she or he might earn. We want to draw a diagram that plots the different combinations of two goods that the person might choose. The goods are hours of leisure time per day (X) and money income (Y), where money income can be spent on consumption goods.

Now suppose that there are 16 hours per day available for leisure or paid work (assuming 8 hours of sleep). Then the hours of work $H = 16 - X$. If a worker's entire income came from paid labor, $Y = wH = w(16 - X)$. But the person might also have some sources of nonlabor income, such as interest income or government transfers. If nonlabor income were \$I per day, then the overall budget constraint would be income per day would be:

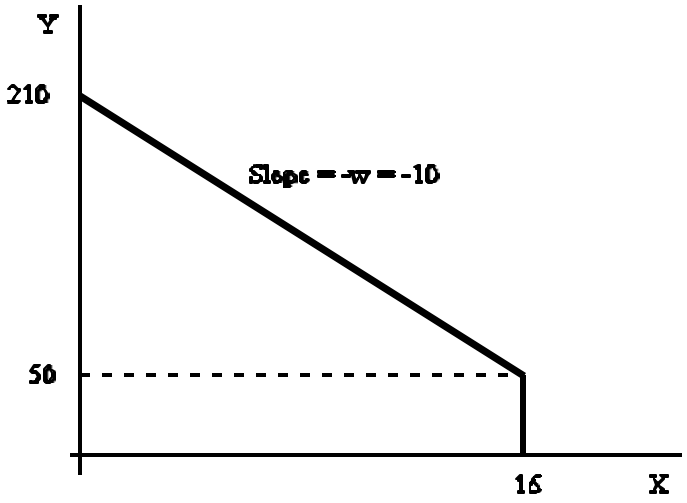
Y = labor income + nonlabor income = $wH + I = w(16 - X) + I$.

Example: Suppose w = \$10/hr and I = \$50/day. Then the budget constraint is

$Y = 10(16 - X) + 50 = 210 - 10X$

This budget constraint is depicted in the following diagram. The upper left endpoint of the budget constraint represents the amount of income the person could earn if they worked all 16 hours (so X = 0). The lower right endpoint represents their income if they didn't work at all (their nonlabor income).

The slope of the budget constraint is -w (negative of the wage rate per hour), and measures the tradeoff between the two goods, leisure hours and income. It is the opportunity cost of one more hour of leisure in terms of foregone income per day.



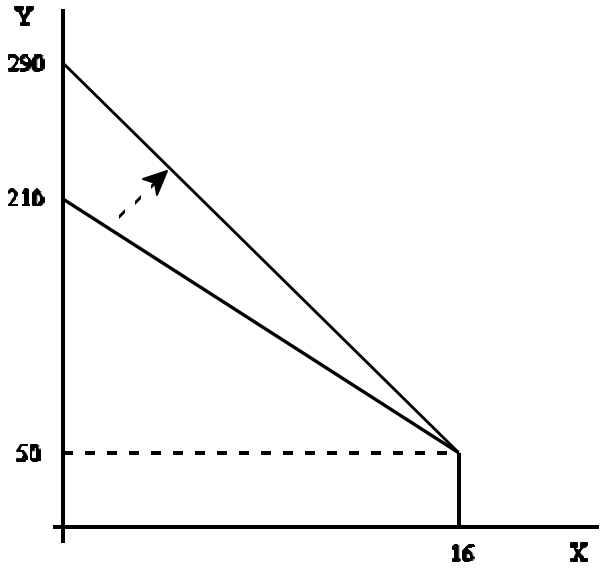
The budget constraint between leisure (X) and consumption spending (Y)

How changes in the wage rate and nonlabor income affect the budget constraint

When the wage rate increases, the budget constraint pivots up (gets steeper); when the wage rate decreases, it pivots down (gets shallower). For example, if the wage in the above example increased to $15 per hour, it would change the budget constraint to

$Y = 15(16 - X) + 50 = 290 - 10X$

The following diagram shows the pivot in the bud\get constraint resulting from this wage increase.



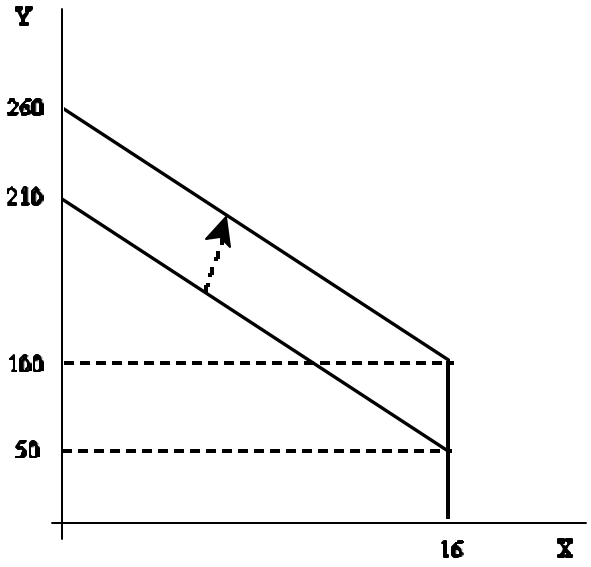
Effect of wage increase on budget constraint

In class I will note that a decrease in the price of consumption goods will also pivot the budget constraint up. For this reason it is most accurate to think of the wage rate in this model as the real wage rate, w/p.

When nonlabor income increases, the budget constraint shifts out (parallel); when nonlabor income decreases, it shifts in (parallel). For example, suppose I increases to $100:

$Y = 10(16 - X) + 100 = 260 - 10X$.

The following diagram shows this change.



Effect of increased nonlabor income on budget constraint

Preferences: introduction to indifference curves

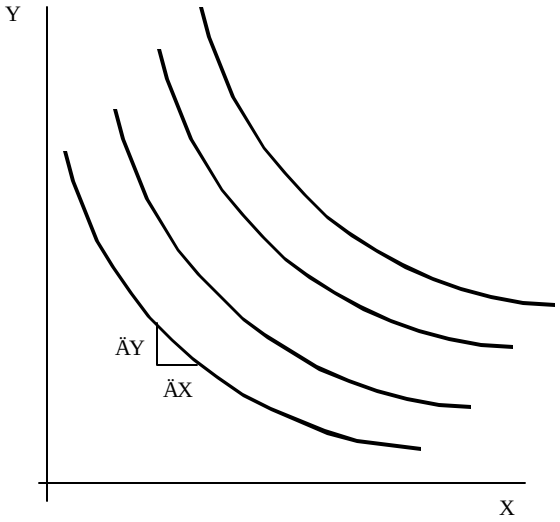
The budget constraint tells us the combinations of leisure time and income (consumption) a person can afford. But which point on the budget constraint should they choose? Answer: We assume they'll choose the combination that *they prefer to any other available.*

To see which combination his is, we need some way of representing the person's preferences in our diagram. We do this with *indifference curves*: An indifference curve connects all the combinations of leisure and income that the person likes equally well (is indifferent between).

We make a few assumptions about preferences and indifference curves:

First assumption about preferences: More is better (a person would rather have more of income or leisure rather than less, all else equal). This implies that

- indifference curves slope down and never cross
- a higher (more preferred) indifference curve lies to the "northeast" of a lower curve
- a wage increase always makes a worker better off, other things equal
- an increase in nonlabor income always makes worker better off, other things equal



Indifference curves

The slope of the indifference curve = ÄY/ÄX tells us how many dollars of income the person would be just willing to give up to get one more hour of leisure time. In other words, it represents the "psychological" dollar cost of an hour of labor time. Economists call this the *marginal rate of substitution*.

Second assumption about preferences: Indifference curves are "bowed in." That means that the more leisure you have, the less income you are willing to give up to get one more hour of leisure. This is also known as diminishing marginal rate of substitution.

Do people really have indifference curves? It's a convenient abstraction. We will focus mostly on the budget constraint and thinking about how people react to shifts in it. The basic idea is that we assume people act rationally and consistently, and that they want more of both income and leisure, if possible.

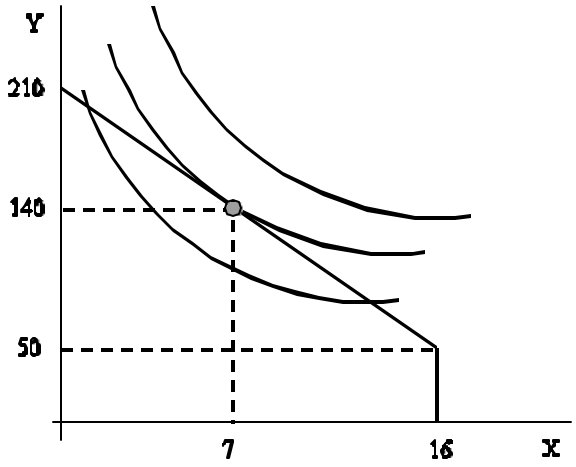
Putting the picture together: the optimal choice of hours and income

The best affordable combination of leisure time and income is the point on the budget constraint that is also the one with the highest possible indifference curve.

This will often occur where the indifference curve is *tangent* to the budget constraint. At that point, the slope of the budget constraint is equal to the slope of the indifference curve,

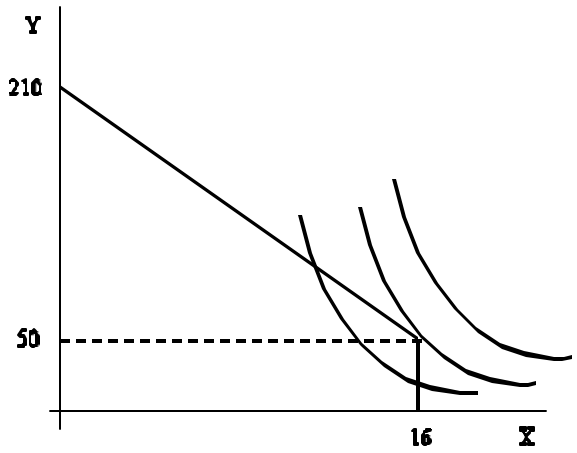
which implies that the psychological cost of one more hour of labor is exactly equal to (offset by) the dollar benefit. That is, the marginal rate of substitution is equal to the wage.

In the following diagram, the individual has chosen 7 hours of leisure per day, implying 9 hours of labor for an income of $Y = 210 - 10(7) = 140$.



Optimal choice at a tangency of indifference curve and budget constraint

A person with different tastes might actually choose not to work at all, which in our diagram would imply that there is no tangency between their indifference curve and their budget constraint. Instead their highest affordable indifference curve is at a *corner solution*, with 0 hours of labor time, as in the following diagram:



Optimal choice at a "corner" when there is no tangency

The effect on work hours of a change in nonlabor income (I)

As we have seen, a change in nonlabor income results in a parallel shift of the budget constraint. How will a person's labor or leisure time respond to this?

Third and final assumption about preferences: Income (consumption) and leisure are both "normal goods". Thus, given an increase in nonlabor income, a person will choose more leisure (less labor), and more total income. I.e., for a parallel shift out they move to the northeast.

Consequently, a person who is working will choose to work fewer hours (take more leisure time) if their nonlabor income increases, and more hours if it decreases. It is as if they use some of their increased nonlabor income to "buy back" a little more leisure time. Given enough nonlabor income, an individual may choose not to work at all (end up at the corner).

The effect of a change in the hourly wage (w)

We have also seen that when the hourly wage changes, the budget constraint pivots (up for a wage increase, down for a decrease). How will this affect hours of work?

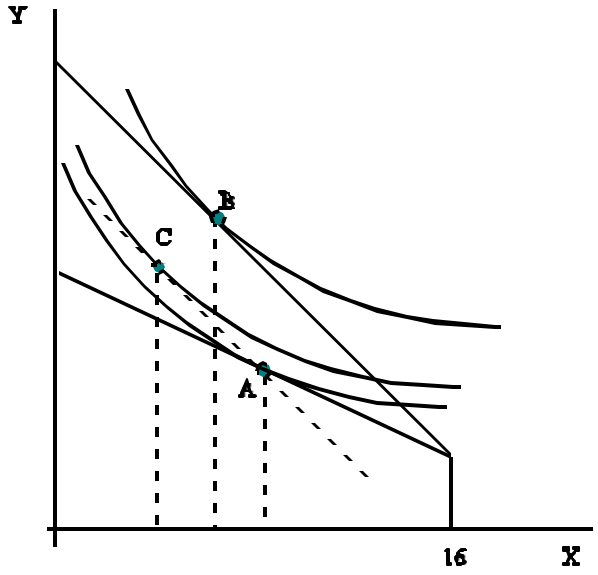
In fact, a wage increase has *substitution and income effects* that work in opposite directions on work hours:

- *Substitution effect:* the wage increase increases the opportunity cost of leisure time, making it more attractive to substitute consumption for leisure (increase work hours)

- *Income effect:* the wage increase allows you (potentially) to have a higher income, which leads to more leisure time (reduce work hours), assuming leisure is a normal good.

Whether a person works more or fewer hours after a wage increase thus depends on which effect is bigger.

Graphically, we can find the substitution and income effects using the following thought experiment. In the following diagram, the overall effect of the wage increase is for the person to move from tangency A to tangency B on their new budget constraint. Our thought experiment divides this effect into two steps. After the wage increase, suppose the person's income were reduced in such a way that they could *just afford* their original consumption point, A, but at the new hourly wage. This would be at the budget constraint indicated by the dotted line. With the dotted budget, they wouldn't stay at A, but would choose the tangent point on the dotted budget, at C. The movement from A to C, which is purely hypothetical, represents the substitution effect because we have "taken away" the change in purchasing power due to the wage increase.



Substitution and income effects of a wage increase

The income effect is captured by the move from C to the ultimate point, B. This is a parallel shift from the dotted budget to the final budget, and it captured the increased purchasing power from the higher wage.

It can be seen that the substitution effect of a wage increase is always negative on leisure (A to C), and therefore positive on labor hours. Conversely, when leisure is a normal good, the income effect of a wage increase is always positive on leisure (C to B), and therefore negative on labor hours.

The upshot is that because these effects work in opposite directions, we cannot predict unambiguously how an individual's labor hours will respond to a wage increase, and therefore we cannot predict whether the labor supply curve of an individual slopes up or down. They might choose to work more hours, as in our diagram, in which case the substitution effect is stronger than the income effect, but for someone else it could go the other way. In fact, the labor supply curve could slope up or down, or be perfectly inelastic (vertical), or it might slope up at low wages and then bend back a high wages, as the income effect takes over. This is known as a backward-bending labor supply curve.

The directions of these effects all reverse themselves when the wage decreases:

	Wage increase	**Wage decrease**
Effect on leisure time:		
Substitution effect	negative	positive
Income effect	positive	negative
Effect on labor hours:		
Substitution effect	positive	negative
Income effect	negative	positive

Other things equal, the income effect tends to be larger the more hours the person was working before the wage change. This is because the change in purchasing power due to a wage change is not very big if the person is not working many hours. Thus in fact the income effect of a wage change is zero for a person who is not currently working. For such a person, the substitution effect is the only effect, and thus a wage increase will, if anything, cause them to

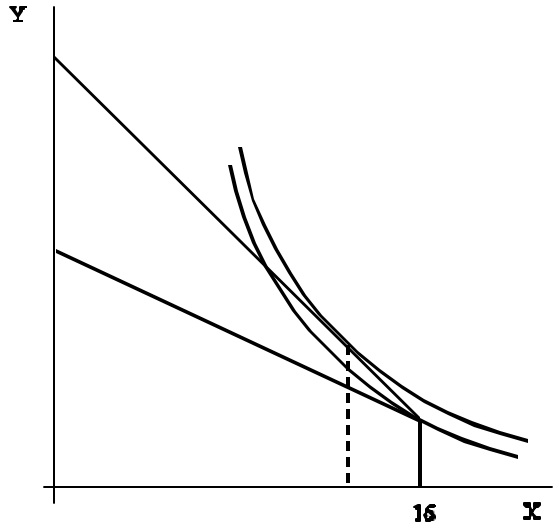
increase their hours.

The participation decision and the reservation wage

So far we have been focusing on the effect of a wage change on the hours of work of a worker. But another interesting aspect of the labor supply decision is whether to work at all. A person who chooses to work positive hours is said to be participating in the labor force. A person who chooses not to work for pay is not participating (not in the labor force, or NILF).

We can use our model of the worker's choice to predict how labor force participation will respond to changing circumstances. In particular, we can say the following:

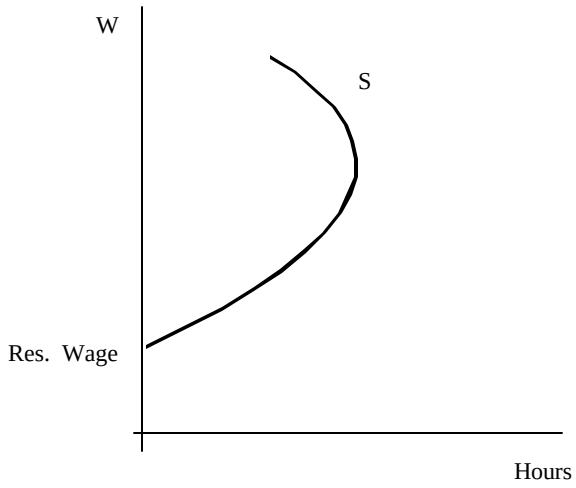
- Increases in nonlabor income, holding the wage constant, make participation less likely. This is because leisure is a normal good, so an increase in nonlabor income leads a person to take more leisure time, and at some point drop out of the labor force altogether.
- Increases in the wage rate an individual could earn will tend to increase participation, as some who would not have participated at a lower wage are drawn into the labor force. This reflects the fact that for nonworkers (NILF), the wage increase only has a substitution effect:



From nonparticipation to participation after wage increase

In class, we will consider how these predictions may help explain the dramatic historical changes in married women's participation rates since around 1940.

The wage rate a which a worker is just at the margin between working and not working for pay is called the *reservation wage*. Putting our analysis together, we can see that the labor supply curve of an individual might look something like the following:



See if you can explain the shape using the concepts of substitution and income effects.

Empirical evidence on individual labor supply

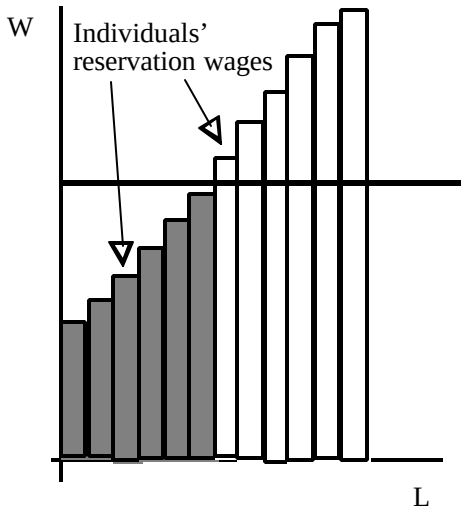
The model of time allocation predicts that an increase in the wage has an ambiguous effect on hours of work (due to offsetting substitution and income effects), but should tend to increase participation rates. Statistical studies of labor supply using data on individuals tends to find the following:

- For adult men, both the income and substitution effects are quite small, perhaps zero. Consequently, men's labor supply is very inelastic.
- For adult women, hours of work tend to be insensitive to the wage as well, and both the effects are small. However, women tend to be more sensitive to wages in their participation decision.

Market supply of labor

The market supply of labor is the sum over all individuals of their hours worked at each wage. In terms of total hours, the market labor supply could slope up or down or be backward bending, just as the individual labor supply curves could be.

If we ignore each person's hours and focus on the *supply of workers* (how many people are in the labor force), then we know that the quantity of workers supplied increases as the wage rises, because more and more workers find the wage rising above their reservation wage:



Market supply: workers with reservation wage below market wage

Note that at any given wage rate, some workers are being paid more than their reservation wage. The difference between what a worker is paid and what she would have to be paid to work (reservation wage) is often called the "rent" earned by the worker.

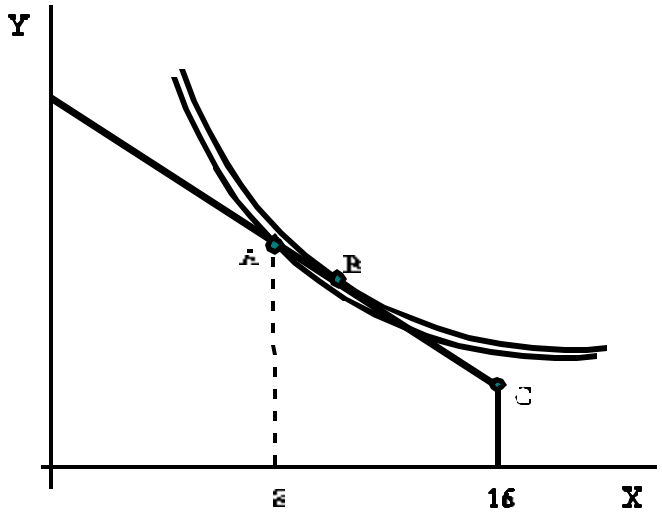
Chapter 4: Labor supply–applications and extensions

In this chapter we look at a couple of applications and extensions of the basic labor supply model. These include the issue of hours restrictions (40-hour work week), work incentive effects of government transfer programs, household labor supply decisions, and recent trends in hours of work.

Hours restrictions and overtime pay

One unrealistic aspect of our model is the assumption that workers are free to choose any number of work hours. In real life, a labor contract often dictates some specific number of hours per day r week (e.g., a 9 to 5 job). Furthermore, federal law restricts hours on many jobs to no more than 40 per week, beyond which overtime must be paid.

We can examine the impact of such restrictions using our basic model. For example, suppose the worker faces a choice between an 8-hour day and no job at all. Then only two points on the budget constraint are available:



Hours restriction prevents optimal choice of hours

Given unconstrained choice of hours, this individual would choose the tangency B, but is forced to choose between A and the corner C. In this case the person prefers A, and ends up working longer hours than she would if unrestricted. But it is also possible that another worker would be forced to work fewer than the desired hours, or might even drop out of the labor force. Therefore, we can conclude that the overall effect of hours restrictions may lead to an increase or decease in total hours and will likely reduce participation.

Overtime pay makes it possible for an individual to work more than 8 hours at a higher wage rate for the hours beyond 8 (say, time-and-a-half). This creates a "kink" in the budget constraint. We will see in class that this may increase or decrease work hours relative to what the person would choose in the absence of overtime provisions.

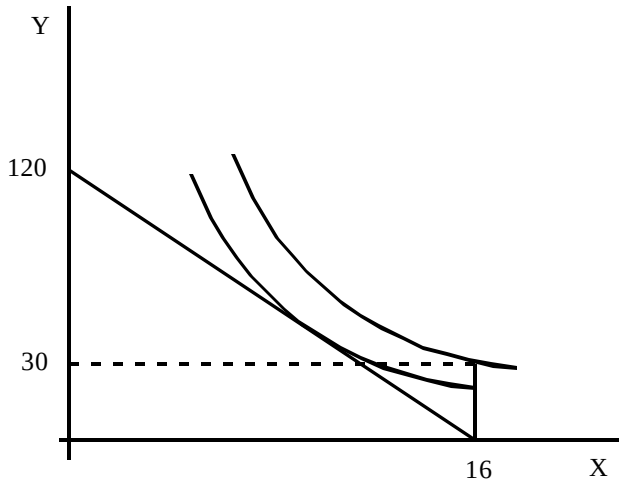
Work incentive effects of alternative government transfer policies

One of the most important policy applications of labor supply theory is in predicting the effects of government programs on labor supply behavior. A particularly interesting example is the effect of government transfer programs (such as welfare) on work behavior. Does the presence of a welfare program affect hours and labor force participation?

A longstanding criticism of welfare programs to help the poor is that they tend to discourage work, and thus lead to dependency and laziness. Indeed, one of the major motivations behind the recent welfare reform (signed into law in 1996, implemented in 1997) was the belief that the older welfare system, AFDC, discouraged welfare recipients from working. The new program sets time limits on welfare payments, and welfare recipients are required to look for and take jobs.

The trick to using our model to understand the work incentive effects of different programs is drawing the budget constraint correctly. Let's go through some different policy alternatives and see what the budget constraint looks like. In each case, I'll assume a worker who can earn $7.50 per hour.

1. First, consider a welfare program that provides a welfare check if a person is not working, but under which any paid work disqualifies the recipient for any payment. This creates a budget constraint with a "spike." Examples of such programs include unemployment insurance, which pays only if you are unemployed, and workers' compensation. Using our example, suppose the government payment were$30 per day if work hours are zero:

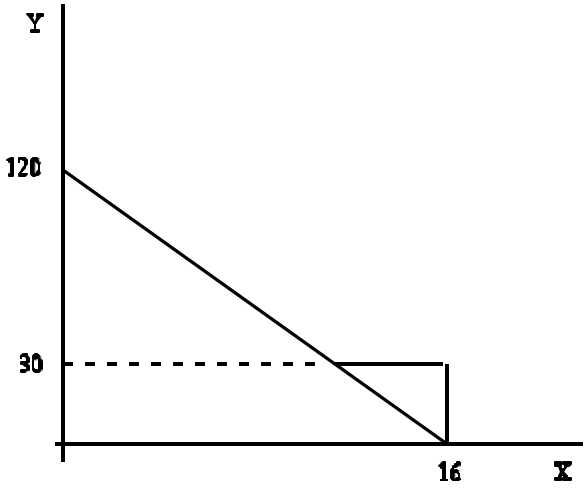


Transfer program with a "spike"

It's not hard to see that this program must reduce the participation of at least some individuals.

2. A somewhat similar policy would to set a *guaranteed minimum income*. For instance,

the government could promise each worker a minimum daily income of $30 by augmenting any earnings that feel below $30. Such a program in effect taxes labor income at 100% below the $30 per day level. The effect on work incentives is identical to the spike.

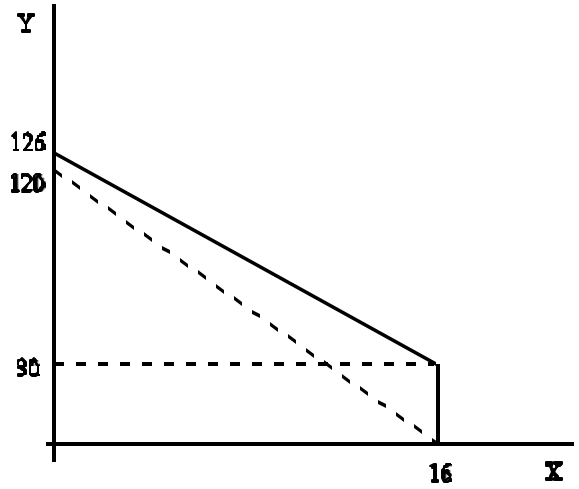


Guaranteed minimum income

A twist on the guaranteed minimum would be to require some minimum amount of work in order to qualify for the payment. As we shall see in class, this creates a funny bump in the budget constraint.

3. *Negative income tax*: The very high marginal tax rate on earned income that is true of some transfer programs led many economists to call for anti-poverty programs that would have smaller work disincentive effects. One important idea along these lines is the negative income tax (NIT). Under a simplified version of the NIT, the income tax has the following formula: $T = tY - S$, where t is the tax rate ($t<1$) and S is a fixed subsidy. Workers with low incomes thus pay a negative tax (receive a subsidy).

Using our example, suppose we have a NIT with $t = 0.2$ (20%) and $S = \$30$ per day. Then labor earnings are $Y = 7.5H$, and then $T = 0.2*7.5H - 30$, so overall disposable income net of taxes is $N = Y - T = 7.5H - 0.2*7.5H + 30 = 6H + 30 = 126 - 6X$.

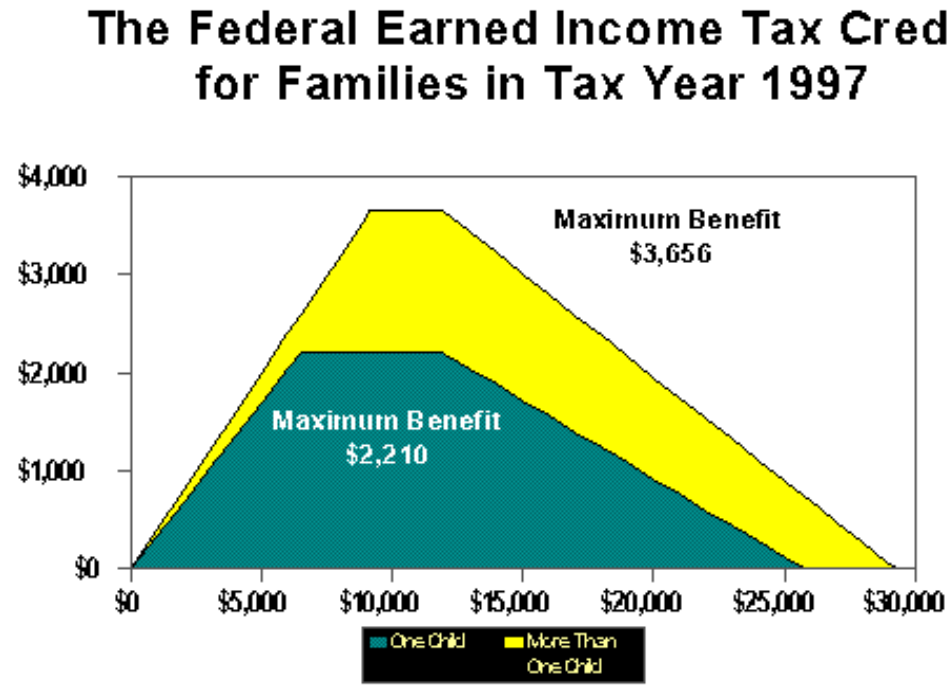


Negative income tax

4. *Earned income tax credit (EITC)*: This is essentially a wage subsidy for poor workers. The subsidy augments a family's income up to a certain level, then is flat, and then phases out. For a family of four, the credit is structured as follows (1997 tax year):

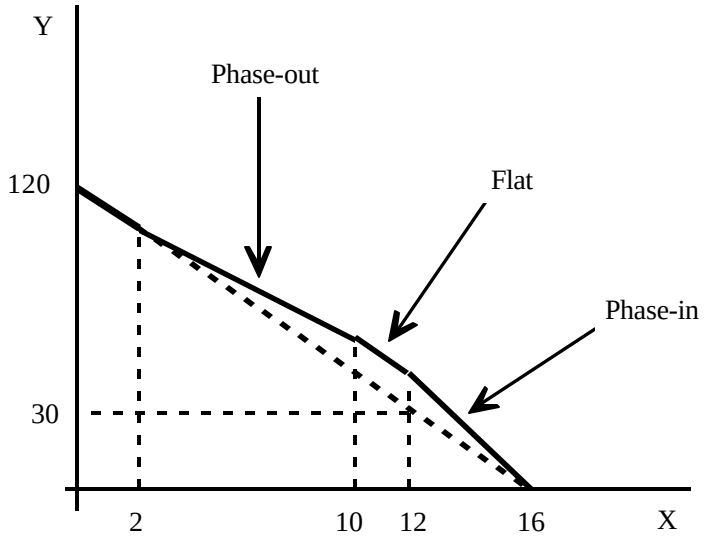
- For workers with annual earnings under $8,900, for each additional dollar of income the worker earns, the government kicks in an additional 40 cents, up to a maximum credit of $3,656.
- Workers earning between $8,900 and $11,610 receive the maximum credit of $3,656.
- For workers earning more than $11,610, the credit is gradually reduced, or phased out, as workers' earnings increase, at a rate of 21 cents per dollar earned.

This diagram shows how the EITC varies with earned income:



These numbers have changed a little bit for the current tax year (the maximum credit for a family of four was up to $3,756 in 1998). The credit is much more generous for families with children than those without. For example, the maximum credit was only $341 in 1998 for a family without a qualifying child.

In terms of our budget constraint, the EITC increases the hourly wage during the "phase-in" part, is like an income subsidy over the flat part, and then is like a tax on the wage for the phase out. A simplified but basically realistic version of the EITC would subsidize each dollar earned by 40% until the worker's earnings hit $30 (at maximum subsidy of $12 per day), maintain that subsidy until earnings reached $45, and then phase it out at 20 cents per dollar earned. Consider our worker with a wage of $7.50. The EITC changes her budget constraint as follows (assuming she is supporting two kids):



Earned income tax credit (EITC)

You can see that the EITC is likely to have a variety of effects, depending on how many hours a person would be working without EITC. Consider individuals who would not be in the labor force in the absence of EITC. The phase-in part of EITC is like a wage increase for them: they are thus more likely to participate. Thus EITC can be expected to increase the proportion working. Next, consider someone who would be working more than four hours per day. For them, the flat part of the EITC is like an increase in nonlabor income (parallel shift): this would induce a reduction in work hours. For someone working even more hours, in the phase-out section, both the substitution and income effects of the policy should tend to reduce work hours.

Thus we can predict that, relative to no EITC, the EITC should increase participation and hours for workers who worked very little, but might well decrease work hours of the already working poor.

Actual effects of transfer programs

What are the actual effects of transfer programs on work behavior? To get a handle on this, one must find data that show how people respond to changes in the programs or benefit levels. One way to do this is to use cross-state differences in support levels. The old AFDC welfare program, for example, varied by about $200 per month between the most and least generous states. Studies that looked at work behavior across states, controlling for other factors, tended to find work effects of different levels of welfare support, but they were rather small. For every $100 gain in monthly benefits, women on welfare worked about 2 hours less per month.

On the other hand, there does appear to have been a strong response to the recent welfare reform, which established work requirements. Research is ongoing to try to identify how much of the increased work of poor people, especially poor single mothers, has been due to the changed incentives under welfare reform versus the strong labor demand conditions in the growing economy.

A further complication is that roughly concurrent with the improving employment picture and welfare reform there have been increases in the generosity of the EITC, which should have drawn more poor workers into the labor force. A recent summary of research on the effects of the EITC suggests that it has played a large role in inducing single parents to go to work, and has had little effect on the hours of poor workers who were already working. This result is hardly surprising, given what we know about labor supply: the substitution and income effects tend to be small for people already in the labor force, but participation (especially of women) is more sensitive to changes in wages and nonlabor income.

All studies that examine the work behavior of poor people must proceed carefully, however, because the official data may underestimate the work hours of the poor. One reason for this is that under AFDC and some other government transfer programs, recipients lose benefits if they report earned income. It may thus be in their interest to earn income "under the table," and they may be reluctant to reveal these earnings to survey takers, who most often are government employees. For this reason, increases in reported labor-force participation that have occurred under welfare reform must be taken with a grain of salt–it may be that some of the increase is due to the fact that recipients are now encouraged to report work activity rather than hide it.

Household production and joint labor supply models

So far we have looked at labor supply decisions as if they were strictly an individual's choice between paid labor time outside the home and leisure time. But of course this simplification is unrealistic: for example, there is also unpaid housework (including child care). Thus we really ought to think of the allocation of time between three activities: paid employment, unpaid housework, and leisure.

In addition, decisions about work time are often made jointly by two or more individuals. Most importantly, in a two-adult household (e.g. married couple), decisions must be made regarding the allocation of time and resources by both parties. Hence we have a problem of "joint labor supply."

Regarding the allocation of time between unpaid household labor and paid employment, we can think of the household combining the household labor time of family members and purchased goods to produce final consumption goods (for instance, a home-cooked meal). The economist Gary Becker thus proposed thinking about a "household production function." Different combinations of these inputs are possible, allowing substitution between household labor and purchased inputs. For instance, family members can do all the housecleaning, or pay someone outside the family to do it. They can purchase labor-saving appliances, or do more work by hand. They can cook at home or buy prepared food, or eat out.

We can think in these terms about the postwar trend toward greater labor-force participation of married women. As women's wages in paid employment have risen in the market, more women have entered the labor force, working for wages. This additional paid work time had to come out of housework time, leisure time, or both. Because we have very little data on how people spend their time at home, there remains considerable controversy about where the hours came from, a question we will examine in class in greater detail.

A particularly important question is how joint labor supply decisions are made in households. For example, in married-couple households, should both adults work for pay? If one is to stay home more hours and do more of the housework or childcare, which should it be?

One way to look at this is to see it as a question of the degree of specialization in household production and labor supply. The theory of comparative advantage, most often applied to international trade, provides some insights here. As you may recall, comparative advantage implies that production can be increased if countries specialize in those activities at which they

are *relatively* more productive. In the context of household labor decisions, the issue involves a comparison of how much each person's labor is worth in the labor market versus home production (housework).

For example, suppose the husband and wife were equally productive at housework, but the husband could earn higher pay in the labor market (this could be due to discrimination or other factors). Comparative advantage would argue for the husband working outside the home and the wife working in the home, so long as the husband was relatively more productive outside the home.

This pattern of specialization could have important feedback effects: as the husband spent more time at his job, his experience and skills would increase, raising his pay, while his skills at housework would atrophy. The reverse might happen to the wife, whose value in the labor market could deteriorate the longer she was out of the workforce. Indeed, researchers find that married women who leave the workforce, usually to take care of small children, suffer a significant loss of earnings potential that takes many years to recover.

In this way, the gender pattern of specialization in the household can be self-reinforcing: the initial pattern of comparative advantage only becomes stronger over time.

The theory of comparative advantage suggests how a household that sought to maximize total household consumption would allocate the labor time of adult family members. This leaves open the important question, however, of whether most households make decisions with this objective. The idea that households act this way can be thought of as an efficiency or household consumption maximizing model. In recent years many economists have challenged this picture of the household. In particular, economists influenced by feminist theory have examined household decisions under the assumption that different members of the household have objectives that may conflict with one another. There may not be a consensus about the distribution of the household's resources, for example.

If conflicting individual objectives are important, household decisions must be reached by some sort of bargaining process. The decisions made will then partly reflect the bargaining power of each family member. In economic models of bargaining, bargaining power is often a function of an individual's "threat point": that is, how well can they do if they withdraw their cooperation? In the case of a married couple, for example, the wife's threat point might be how well she could do for herself if she chose not to specialize in household production but instead

earned and spent her own income. Or it might be how she could do on her own if separated from her husband.

Although the details are beyond the scope of this course, issues of bargaining power could have important implications for time allocation decisions. For example, a woman who was concerned about her bargaining power in the relationship might choose to work outside the home in order to have a better threat point, even if comparative advantage would suggest that she specialize in housework.

In addition to bargaining considerations, custom and social norms play an important in household decision making. The social appropriateness of married women working outside the home, for example, can vary dramatically across cultures and over time within a culture.

Chapter 5: Labor demand

Whereas individuals and households supply labor time to the market, the demand for labor is more often by organizations, including private business firms, government, and nonprofit organizations (of course, households also employ workers). In this chapter we will focus on the demand for labor by profit-maximizing business firms. To the extent that other types of organizations hire labor with the goal of maximizing some objective that uses labor as an input, many of the considerations we examine here will apply to those organizations as well.

One important objective of this chapter is to examine the determinants of the elasticity of labor demand. As we shall see, the elasticity or wage-responsiveness of demand plays a very important role in predicting the effects of the minimum wage on employment, and also has very interesting implications for understanding the impact of globalization on American workers.

Profit maximization and cost minimization

The theory of labor demand begins with the standard microeconomic assumptions about the behavior of private business firms. Namely, we assume that businesses produce and sell goods or services in order to maximize their economic profit. In order to maximize profits, firms must choose not only the best output level and price but also the best amount of each input. Whereas in you core microeconomics courses you probably focused on the firm's decisions in the output market, we emphasize the choice of inputs (especially labor) here.

Labor demand in the short run: the marginal product theory

We begin with the short run. Recall that in the short run, some inputs are fixed and some are variable. We assume that labor is a variable input, whereas capital is fixed in the short run. You can think of this as a firm with some fixed capacity (plant and equipment, or office space) trying to decide how many workers to employ given that capacity and the market for its product.

To determine how many workers the firm should hire, we can employ the usual marginalist logic: should the firm hire one more worker? To answer this, the firm should compare the benefits that worker brings to the firm in the form of increased production and revenue against the cost of paying the worker.

The basic rule for profit maximizing employment is that the firm should keep adding more units of labor input until the additional revenue generated by its product just equals its added cost. In economic jargon, the firm chooses employment where marginal revenue product of labor is equal to the marginal expense of labor.

- *Marginal revenue product* of labor (MRP_L): the additional revenue from selling the product of one more unit of labor
- *Marginal expense* of labor (ME_L): the additional cost associated with employing and paying one more unit of labor.
- Profit-maximizing employment level requires $MRP_L = ME_L$

Let's figure out what MRP and ME actually are. First, what does one more worker add to revenue (MRP)? It's simply the number of units she adds to production times the value of each additional unit added. The addition to production is the marginal product of labor (MP_L)--specifically, the increase in output for a small increase in one input, holding all other inputs fixed:

$$MP_L = \frac{\Delta Q}{\Delta L} \quad (other\ inputs\ fixed)$$

Generally, when there are fixed inputs, we assume there is *decreasing marginal product* as additional units of labor are added. (Why?)

Given the marginal product, to obtain the marginal revenue product we need to multiply MP by the marginal revenue (MR) obtained from selling one more unit of product. Therefore,

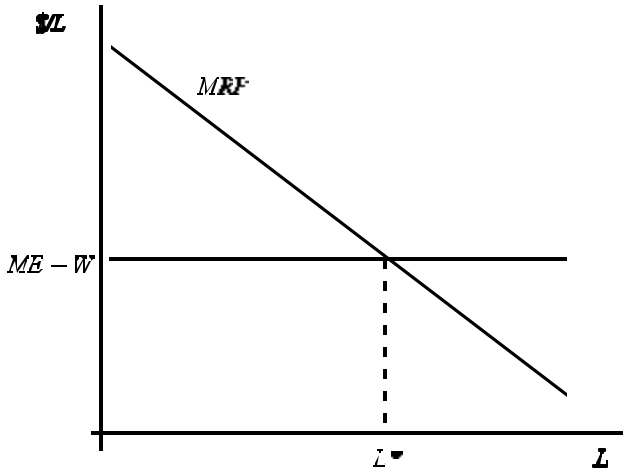
$$MRP_L = \frac{\Delta R}{\Delta L} = \frac{\Delta R}{\Delta Q} \cdot \frac{\Delta Q}{\Delta L} = MP_L \cdot MR$$

If the product (output) market is perfectly competitive, then MR = P, so

$$MRP_L = MP_L \cdot MR = MP_L \cdot P$$

Now that we've derived the marginal revenue product, let's consider the cost or expense side. The marginal expense of labor is simply the cost of one more unit of labor. In a competitive labor market, ME_L = w, the wage paid per unit of labor.

Now we're ready to put it all together. In order to maximize profits, the firm must choose its labor input where $MRP_L = ME_L$. Graphically,



Optimal labor demand where w = MRP

See if you can explain why the firm would not be maximizing profits if it operated with an employment level above or below L*.

If both product and labor markets are competitive, profit maximization requires:

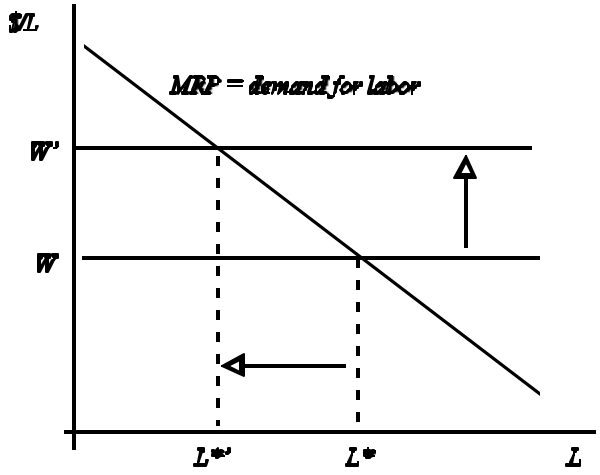
$$MRP_L = MP_L \cdot P = ME_L = w \quad or \quad MP_L = \frac{w}{P}$$

That is, the marginal product of labor must equal the "real product wage" (the money wage divided by the price of the product). The firm has chosen the right employment level (is maximizing its profit) if the last worker added produces an amount of product that exactly pays for the value of her wage measured in terms of the product.

Do firms actually choose their labor input in this way? What if they don't know how to maximize profits, or don't try? One answer to these skeptical questions is that although there may be many factors affecting employment decisions by firms, profitability (maximizing shareholder value) is likely to play an important role. Secondly, even if firms make their decisions by habit or using simple rules of thumb, competitive pressures tend to weed out rules of thumb that do not enhance profitability.

The demand curve for labor in the short run

Now we can use our model to derive the firm's demand for labor in a competitive labor market by varying the wage rate in the market and seeing how the firm responds. In the following diagram, we can see that as the wage rises or falls, the firm chooses employment along the marginal revenue product curve. In other words, *the marginal revenue product curve is the firm's demand curve for labor.* It slopes downward, as expected of a demand curve.



Effect of wage increase on labor demanded

Let's work through a simple numerical example. Consider a garment shop, where daily

production as a function of labor (employment) is $Q = 20L - 0.05L^2$, workers are paid a daily wage $w = \$80$, and the garment is sold wholesale for $P = \$20$.

Then using a little calculus, $MP_L = dQ/dL = 20 - 0.1L$. We can set this equal to w/P: $MP_L = 20 - 0.1L = w/P = 4 \Rightarrow L^* = 160$. Thus optimal employment is 160 workers.

What would the demand curve for labor be? We use the same formula, but allow the wage to be at any value by keeping it as a variable, w. The formula then requires MP = w/P, or $20 - 0.1L = w/20$. Solving for L, we obtain $L = 200 - w/2$.

This gives us the demand curve for labor for a single business firm. Of course, to obtain the market demand we would want to add up the demand by all the firms in the labor market. For instance, suppose there were 100 garment firms just like this one in the market for garment workers. Then the market demand for garment labor would be $L_{market} = 100L = 20{,}000 - 50w$.

Shifts in the labor demand curve

Because the short-run labor demand curve is the marginal revenue product curve, anything that shifts the MRP curve necessarily shifts the demand curve. Since MRP = MP*MR, there are two obvious sources of shifts in MRP. First, anything that shifts the marginal revenue will shift MRP. Essentially, this means that if the demand for the firm's product, or the price of that product, increases, demand for labor will increase. In this sense we can speak of the demand for labor as derived demand, because it is derived from the demand for the product that the labor produces.

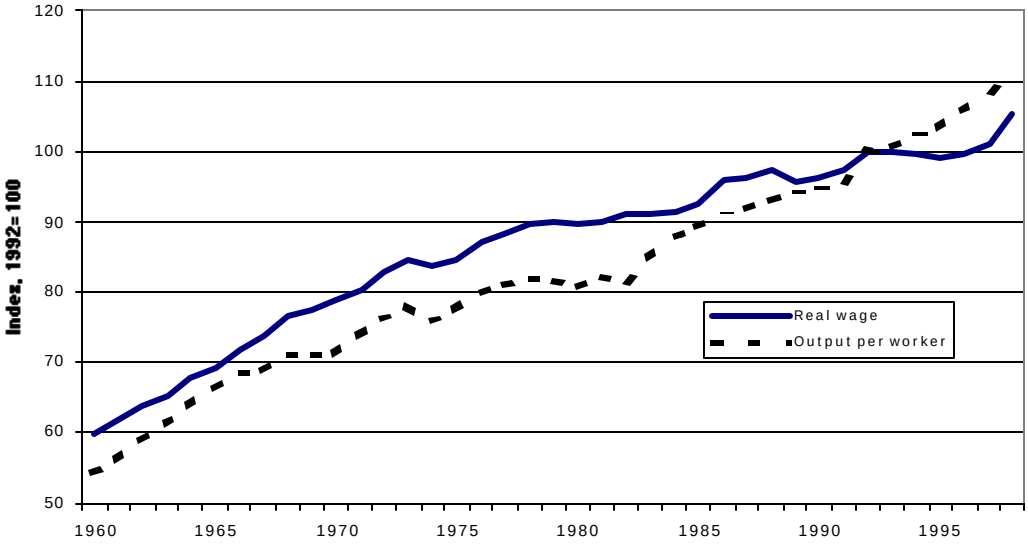
Second, anything that shifts the MP curve will also shift labor demand. The MP of labor can shift for two reasons: (1) technological change that alters the productivity of labor; (2) changes in the levels of other inputs, such as capital. For instance, if a firm increases its capacity by adding to its productive capital, this will generally increase the productivity of labor at each employment level, shifting the MP curve up. This in turn increases the demand for labor at any wage.

Wages and productivity

The marginal product theory predicts a strong link between real hourly wages and worker output per labor hour. Thus one of the strongest explanations of the slowdown in average wage growth since the 1970s is that the growth of productivity (output per labor hour) has slowed

during the same period.

The following figure shows indexes of real hourly wages and labor productivity since 1960, for the U.S. private business sector. The slowdown in the growth of wages after around 1973 is evident. It is also clear that productivity grew more slowly during the period 1975-1990 than it did during 1960-1975. Since 1990, however, productivity growth has picked up substantially, although it is still growing more slowly than it did during the two decades following World War II. Wage growth has lagged the surge in productivity during the 1990s, although there is evidence that wages are beginning to rise. Whether the current growth in productivity represents a temporary phenomenon or a return to rapid economic growth driven by high-technology (the "new economy") is a subject of considerable debate among economists.



Real wages and labor productivity, 1960-1998

Ethical interpretations of the marginal productivity theory

The competitive theory of labor demand implies that workers in the market get paid the value of their marginal product. It turns out that by a similar logic, other factors of production, such as owners of capital, will be compensated at the value of their marginal product as well. The great American economist John Bates Clark, who co-discovered the marginal product theory, argued that the market wage is the just or ethical wage for precisely this reason. Because marginal

product measures a worker's contribution to production, so the competitive wage reflects that contribution too. In a book published around 1890, he wrote:

> If each productive function is paid according to the amount of its product, then each man gets what he himself produces. If he works, he gets what he creates for working; if he also provides capital, he gets what his capital produces; and if, further, he renders service by coordinating labor and capital [entrepreneur], he gets the product that can be separately traced to that function. Only in one of these ways can a man produce anything.... If wages, interest and profits, in themselves considered, are fixed according to a sound principle, then the different classes of men who combine their forces in industry have no grievances against each other.

In other words, in Clark's view, the claims of trade unionists or socialists like Karl Marx that the free-market wage is unjust or that workers are paid an exploitative wage is simply incorrect.

We will discuss Clark's interpretation in class. Do you agree with Clark's view? What flaws, if any, do you see in his argument? As you think about it, bear in mind that one need not accept Clark's ethical interpretation to accept his economic theory of labor demand.

Labor demand in the long run

In the long run, the firm is free to vary all inputs. It can choose optimal levels of employment and capital stock. The marginal product rule is still valid in the long run: the firm wants to choose each input in such a way that the input's MP is equal to the ratio of the input price to the output price.

We saw that in the short run, an increase in the wage leads to a reduction in the quantity of labor demanded. How does a change in the wage rate affect the quantity of labor demanded in the long run?

In the long run, the wage change has scale and substitution effects.

- *Scale effect:* when the wage rate rises, production costs rise, and the firm tends to produce less. This tends to reduce the quantity of labor demanded.
- *Substitution effect:* When the wage rises, the firm might substitute other inputs for labor, because labor is now relatively more costly.

Both effects tend to be negative when the wage rises, so the long-run demand for labor, like the short-run, is downward sloping. (Both effects are positive when the wage falls, of course.) Because the substitution effect is only relevant in the long run, when the firm is free to substitute other inputs for labor, the demand for labor is likely to be more sensitive to wage changes in the

long run than it is in the short run. The sensitivity of quantity demanded to price (wage) changes is of course the elasticity of demand.

Elasticity of labor demand

The sensitivity of labor demand to changes in the wage rate–in other words, the elasticity of labor demand–is a critical issue in labor economics. Demand elasticity is important in understanding the impact of globalization on American workers as well as the employment effects of the minimum wage, as we shall see.

Recall that the definition of elasticity (as applied to labor markets) is the percentage change in quantity of labor demanded for a percentage change in the wage, or:

$$\eta = \frac{\%\Delta E}{\%\Delta w} = \frac{d\ln E}{d\ln w} \qquad (\eta \textit{ is the Greek letter "eta"})$$

where *ln* is the natural log. Why does the formula with logarithms give us elasticity? Recall from calculus that dlnE = dE/E = %ÄE (after multiplying by 100). Similarly, dlnw = dw/w = %Äw.

Since the demand curve slopes down, the elasticity is a negative number. The larger ç is in magnitude, the more elastic is demand (demand curve is flatter).

What factors determine the elasticity of labor demand? The response of firms to a change in the wage depends on the output and substitution effects. Factors that make these effects larger tend to make demand more elastic. We can identify the following important determinants of labor demand elasticity:

Operating through the output effect:

- *The demand for a particular class of labor tends to be more elastic if demand for the product is more price elastic*. As the wage rises, the increase in unit cost leads to an increase in the output price, which leads to a drop in sales (output effect), which is greater the more sensitive product demand is to price changes. This is one reason that labor demand is likely to be more elastic in the long run than the short run, because product demand itself is usually more elastic in the long run.
- *The demand for a particular class of labor tends to be more elastic if the cost of employing that class of labor is a larger share of total costs*. The impact of a wage change on unit costs and therefore price and sales will be a function of how important

labor is as an input. Other things equal, labor demand should be less elastic in a highly capital-intensive industry, such as oil refining, than it is in a highly labor intensive industry, such as garment manufacturing or fast food sales.

Operating through the substitution effect:

- *The demand for a particular class of labor tends to be more elastic if other inputs can be easily substituted for that class of labor.* If substitution is easier, the substitution effect will be larger.
- *The demand for a particular class of labor tends to be more elastic if the supply of other inputs is more elastic*. The operation of the substitution effect would be limited if the other input prices rose as the firms tried to purchase more of them. But if their supply is elastic, then their price is not much affected by the increased demand, and so substitution continues.

Because there is no substitution of inputs in the short run, the labor demand response to a wage change should tend to become more elastic given a longer period of adjustment.

What is the elasticity of labor demand in the real world? The best estimates tend to find demand elasticities around -1, but as we shall see, some recent evidence about the employment effects of the minimum wage suggest that labor demand may be much less elastic than that.

Cross-price effects on labor demand

Another important aspect of labor demand in the long run is the effect of changes in the price of one input on the demand for another. For example, how would a decrease in the price of computers affect the demand for different types of labor? These effects are known as *cross-price effects*.

Cross-price effects can depend on whether the inputs in question are complements or substitutes in production.

- Two inputs are considered *complements in production* if the increase in the use of one input tends to increase the use of the other.
- Two inputs are considered *substitutes in production* if the increase in the use of one input tends to decrease the use of the other.

We can now see how a change in one input price might affect the demand for another input. When one input price goes up, the effect on the demand for other inputs includes both scale and

substitution effects. The scale effect is negative for all inputs, because higher costs cause the firm to cut back on production, which tends to reduce demand for every input.

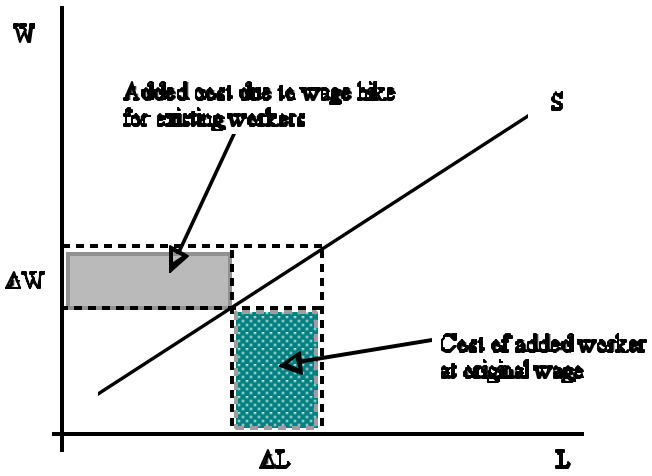
But the direction of the substitution effect depends on whether the inputs are complements or substitutes. To see how, let's consider the case of three inputs at a bill processing facility: clerical workers, computers, and computer programmers. What is the effect on demand for the two types of labor when the price of computers falls?

- Clerical workers: The scale effect is positive. Overall costs are lower, so the firm expands output and tends to need more inputs. But the substitution effect is probably negative: the firm can substitute computer power for some of the tasks that clerical workers do, such as hand-processing bills.
- Programmers: The scale effect again is positive. Here the substitution effect is also likely to be positive, because programmers and computers are complements in production. They go together. The decline in the cost of computing leads to a substitution of BOTH computers and programmers in place of clerical workers.

Monopsony

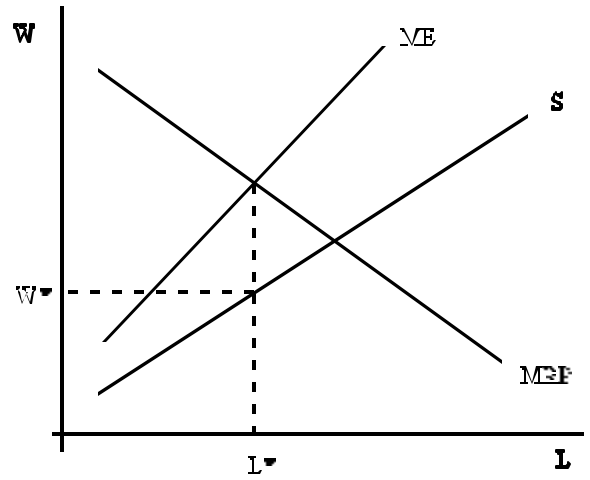
Monopsony means a single buyer. It is the analogue to monopoly, but on the buyer side. As with monopoly, there are degrees of monopsony power. If there is truly only one employer of labor, we have pure monopsony. But there may be some degree of monopsony power even if there are more than one employer. The crucial aspect of monopsony is that the employer does not take the wage as fixed by the market. In other words, *if the employer were to cut the wage, it would not lose all its employees*, as a competitive employer would.

We can assume that a monopsonist still wants to maximize profits, and therefore would still want to choose L such that MRP = ME. But where the monopsony differs from competition is that in general, $ME_L > w$. Why? In order to hire one more worker, the employer must raise the wage a little for all workers. (Assumption: The employer must pay all similar workers the same wage—there is no price/wage discrimination.) Thus the additional cost of hiring one more worker is not just the new worker's wage, but also the added expense of paying all the employees a little more. In the following diagram, S is the supply of labor facing the monopsony firm.



Cost of one more worker to a monopsonist

Because the ME > w, the ME curve for a monopsony is a curve above the supply curve (as a monopoly's marginal revenue is below the demand curve).



Optimal employment and wage for monopsony

The monopsony firm maximizes profit by choosing L* where MRP = ME. The wage rate it charges is the minimum wage necessary to employ L* workers, which is given by the wage on the supply curve W*. Note that if the market were competitive, the firm would operate where MRP (the competitive demand curve) crossed the supply curve. Thus we can conclude that compared with the competitive market, a monopsony

- employs fewer workers and
- pays a lower wage.

Returning to our numerical example, let's derive the optimal wage and employment for a monopsony. Recall that $MP_L = 20 - 0.1$, and $P = 20$. Then $MRP = P*MP = 400 - 2L$. Now suppose the firm faces a labor supply curve given by $L = w + 100$. To obtain the ME curve, we first solve the supply curve for $w = L - 100$. The expense of labor is $E = wL = (L - 100)L = L^2 - 100L$. Then $ME = dE/dL = 2L - 100$. Setting ME = MRP we have $2L - 100 = 400 - 2L \Rightarrow L^* = 125$. Plugging this back into the supply curve, we have $w^* = 25$.

How relevant is monopsony to the real world? Are there any monopsonies out there? There are a few fairly plausible examples, including company towns, such as some of the remote mining towns of the 1800s and early 1900s, and professional sports before free agency. But economists have generally been skeptical that much monopsony power exists in most labor markets because it appears that in most places, workers do have a number of employers to choose from.

Still, there may be more subtle sources of monopsony power. The essence of monopsony power is that the employer can lower the wage without expecting to lose all the employees. This is bound to be the case to the extent that some employees face search or mobility costs to changing jobs. This may arise because some workers have developed firm-specific skills that cannot easily be transferred to new employers, or if information about alternative jobs is imperfect. One indication that job search is costly to workers is that a large percentage of workers searching for a new job will accept the first job offer they get, rather than continue to look for a better offer.

Employment effects of the minimum wage

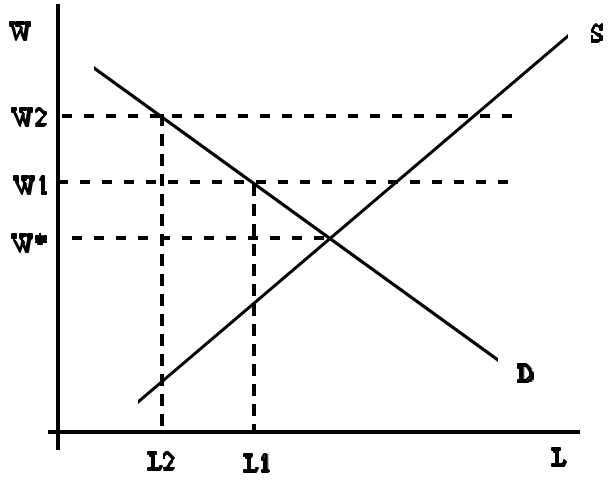
The federal minimum wage was established during the 1930s as part of the New Deal

response to the Great Depression. Over the years the nominal level of the minimum wage has been increased periodically by Congress, and the percentage of workers covered by the law (subject to the minimum) has increased as well. In class we will examine how the real value of the minimum wage has changed over time, as well as its level relative to the median wage. In addition to the federal minimum wage, many states (including California) have their own minimum wage set above the U.S. minimum. Recently, a number of cities (including San Jose) have passed living wage ordinances that set a still higher minimum wage, although in many cases the minimum applies only to city workers and firms that contract with the city.

The motivation for adopting a minimum wage is clear enough: it aims to place a floor under the income of the very poorest workers. Unlike some types of transfer programs, it does so without removing the incentive to work. Yet mainstream economists have been very critical of the minimum wage over the years. The most frequent criticism is that the minimum wage, if binding, is like any binding price floor, creating excess supply, in this case taking the form of involuntary unemployment. Because it is the low-wage workers for whom the minimum is binding, they are also the ones who will suffer the unemployment. Employed workers are helped by the minimum, but others lose their jobs.

In this section I examine some theoretical predictions about the effect of the legal minimum wage on employment under competition and monopsony. Theory suggests that in competitive labor markets, at least, there should be a tradeoff between higher wages and reduced employment caused by the minimum wage. But what are the terms of the tradeoff? In class we will also read and discuss some research attempting to estimate the actual employment effects of the minimum wage in the real world.

The standard competitive model of the labor market predicts unambiguously that an increase in the minimum wage will decrease employment. This can be seen in the following diagram. When the minimum wage is binding (above W*), an increase from W1 to W2 causes a reduction in employment from L1 to L2:



Employment effect of hike in minimum wage, perfect competition

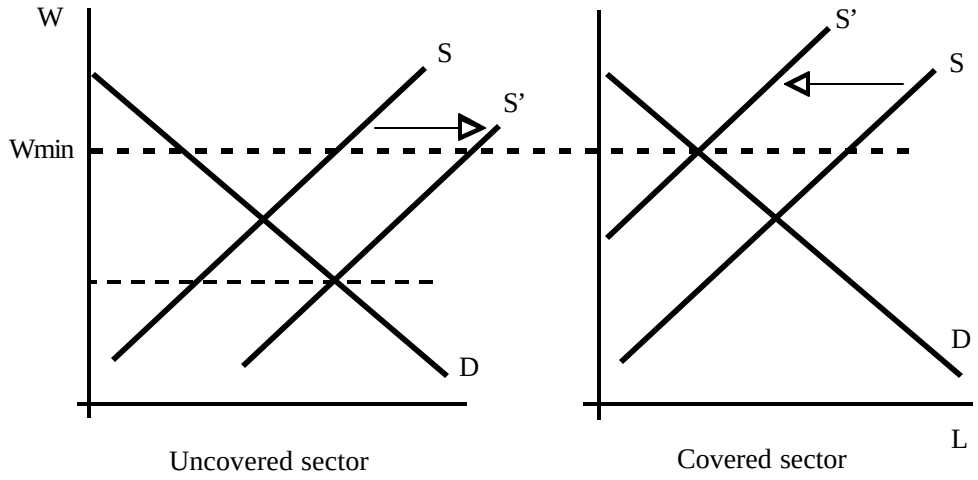
Clearly the impact on employment of a binding minimum wage depends on the elasticity of demand. If the demand elasticity is -1, for example, it implies that a 10% hike in the minimum wage will reduce the employment of minimum-wage workers by 10%. Because labor economists have typically found demand elasticities on the order of -1, many have predicted that the minimum wage should have fairly sizeable effects on employment.

Covered and uncovered sectors

In its early years, the federal minimum wage law did not apply to a large percentage of workers: workers in agriculture, for example, were not covered. Coverage has been extended over the years to most workers. However, many living wage laws apply only to workers employed by a city or firms with city contracts. In these cases we can speak of covered and uncovered sectors. When some firms are covered and others are not, the employment effects of the minimum wage are complex.

In the following diagram, we assume for simplicity that the equilibrium wage in the absence of the minimum wage would be the same in both sectors. But after the minimum is introduced in the covered sector, its wage rises. This causes unemployment, forcing some workers in the covered sector to pursue jobs in the uncovered sector. As a consequence, supply shifts from the

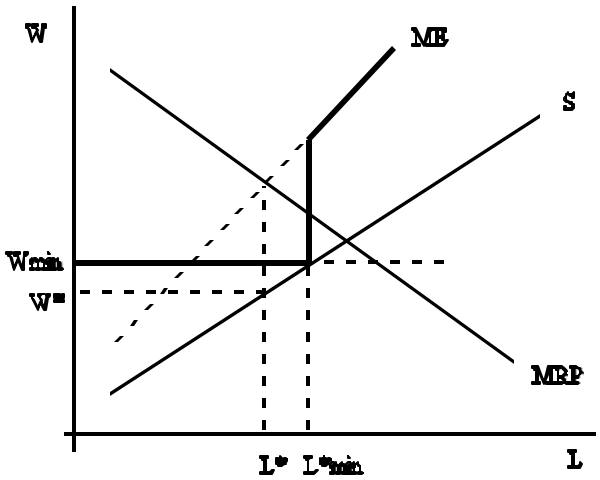
covered to uncovered sector. In the diagram, this has continued until the wage in the covered sector had risen to the minimum, eliminating the unemployment. The workers in the uncovered sector, however, find themselves earning less than they did without the minimum wage law, due to the crowding of new workers into their market.



Minimum wage with covered and uncovered sectors

Monopsony and the minimum wage

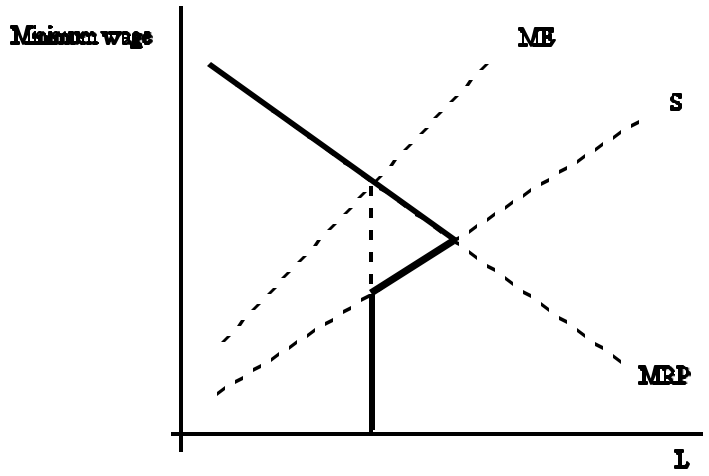
The effect of an increase in the minimum wage can be very different for a monopsony. The legal minimum essentially makes the supply of labor to the firm perfectly elastic at or below the mandated minimum wage: the firm is not permitted to cut the wage below the mandated level. Thus the ME_L curve is flat up to the point where the minimum wage crosses the labor supply curve, and then jumps up to the unregulated ME curve at a higher wage. In the figure below, the ME is the heavy curve. Note that ME crosses the MRP curve on the vertical portion of the ME. The monopsony's optimal employment with the minimum wage is shown at Lmin, which is higher than the L* it would choose if unregulated.



Monopsony employment with minimum wage

Why might the monopsony employ more workers when forced to pay a higher wage? The reason is that the monopsony firm benefits by reducing the wage, but only at the cost of losing some employees. But when the firm is unable to reduce the wage as far as it would like, it has no reason to reduce employment that much.

How will the monopsony employment respond to an increase in the minimum wage? If you imagine a small increase in Wmin in the above diagram, you can see that employment will increase. That is, it will until the minimum reaches the point where S crosses MRP. At that point, further increases in the minimum lead to reductions in employment along the MRP curve, just as they would for a competitive firm. So the relationship between the minimum wage and employment by a monopsony is actually quite complex. The minimum wage is not binding on a monopsony if it is set below the optimal W*. As the minimum rises above W*, employment at first increases and then eventually decreases. This is shown in the following diagram.



Monopsony employment at different values of minimum wage

The significance of the monopsony analysis is that it describes conditions under which an increase in the minimum wage could actually increase employment, up to a point. In class we will discuss some other theories that might help us understand why the minimum wage appears to have relatively little impact on employment according to some economists.

Chapter 6: The determination of earnings

We now have taken a close look at the supply and demand sides of the labor market. Putting them together in a competitive market model gives us a theory of equilibrium wages, as modified by possible monopsony power, unions, government regulation, etc. But the theory remains very abstract, for we have not really addressed a crucial issue: explaining the large observed differences in pay across different workers, and different jobs or occupations.

Looking at the supply and demand diagram, we can see that wages will tend to be higher in a job market if there is relatively less supply and/or relatively more demand. Demand tends to be greater if the workers are more productive (MP theory). But that doesn't explain how they got to be that way. Maybe some workers are just born with more native talent. But there are also things people can do to enhance their productivity: education, training, on-the-job learning, etc. Adam Smith argued that the variation in natural talent is actually rather small, and that most differences in abilities and occupations are learned. Whether or not he was right, can we say more about the patterns of pay differentials we observe in the labor market?

In this chapter, we will begin with a simple competitive market model of pay differentials known as the theory of compensating wage differentials, or equal net advantage. This is the subject of the reading from Adam Smith (Wealth of Nations, chapter 10). We will then explore two alternative models of the role of education in the determination of earnings: the human capital and labor-market signaling models. Finally, we will briefly examine some social and institutional factors that affect pay, including labor unions and discrimination.

The theory of compensating wage differentials

Why do some jobs or occupations pay more than others? The theory of compensating wage differentials explains differences in pay across different jobs as reflecting differences in the *nonwage costs and benefits* associated with those jobs. Jobs with nonwage characteristics that workers don't like will require higher pay to attract workers in the market. Jobs with nonwage characteristics that make the job relatively attractive to workers will be able to pay workers relatively less. In this sense the competitive labor market will *compensate* workers for undesirable nonwage job characteristics.

Note: This is *not* a matter of saying that workers *should or should not* be compensated for undesirable work, nor is it a matter of saying that employers compensate workers for undesirable work because they think it is the right thing to do. Rather, the model merely predicts that the compensation *will* occur in the market equilibrium.

Let's now derive the basic theory. Three key assumptions underlie the model of compensating differentials:

- Rational choice by workers: each worker seeks the best job, all things considered.
- Complete information: workers know which jobs are available at what rates of pay.
- Mobility: workers are free to move to a preferred job.

If these assumptions hold, then we can state the basic result:

The principle of equal net advantage. In the competitive equilibrium of the labor market, wage differentials between jobs are such that the jobs have equal net advantage for the marginal worker: the sum of wage (pecuniary) and nonwage (nonpecuniary) net benefits are equal across jobs. Thus jobs that are less desirable should pay higher wages, other things equal.

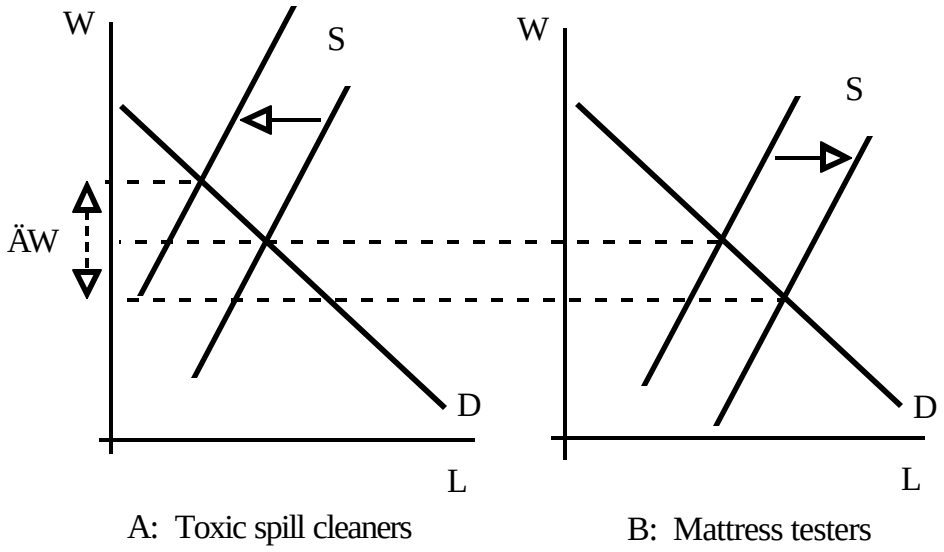
The proof of this result is fairly intuitive. Given our assumptions, over time a worker is free to switch jobs if the overall combination of pay and non-wage net benefits is better on another job. The labor market is not in equilibrium until workers stop switching jobs. Thus in the competitive equilibrium, each worker must hold a job with a wage and nonwage package that she views as at least as good as any alternative.

For some workers, this will mean that they definitely prefer their job to any other job available in the market. These workers are "inframarginal," in the sense that they are not tempted to switch. Some workers, on the other hand, will just be indifferent (at the margin) about switching jobs. These are the so-called marginal workers--marginal in the sense of "at the

margin," not in the sense of bad or dispensable. Net advantage is equal for the marginal worker.

Note the specific sense in which each worker must prefer her job to every other job. Consider someone with an unpleasant job, such as Adam Smith's public executioner. This individual might prefer to be a brain surgeon, and earn a brain surgeon's salary. But the amount of time, effort, and expense that would be required for this person to become a brain would be exorbitant. It is in this sense that they actually prefer being an executioner to being a brain surgeon, given all that the latter entails.

A simple supply-and-demand analysis illustrates the principle. Consider two jobs that require equal skill. Job A is highly unpleasant (say, toxic spill cleaner), and job B is very pleasurable (mattress tester). In the following diagrams, the wages in these jobs are initially equal. But *this cannot be a market equilibrium*, because some of the toxic spill cleaners will want to be mattress testers if they can earn the same wage. So supply is reduced in market A and increased in market B, as some workers change jobs. Given the demand, this raises the wage in A and reduces the wage in B. The process continues until the wage differential (ÄW) between the two jobs is sufficiently great that no more toxic spill cleaners want to switch to be mattress testers.



What kinds of job characteristics can give rise to compensating differentials? Essentially,

anything that workers care about. Adam Smith listed five: (1) agreeableness, (2) cost of learning, (3) constancy of employment, (4) trust required, and (5) probability of success. Make sure you understand his argument for each.

If the market worked this way in most cases, note that workers in dangerous or risky jobs would be compensated for their work hazards in the form of higher wages. Free-market economists cite this outcome as an argument against government regulation of workplace safety, such as OSHA. In their view, the market must provide adequate compensation for work hazards, or these workers would simply switch to safer jobs.

The competitive model of compensating differentials or equal net advantage has various limitations. First, even if the necessary conditions hold, the theory only tells us that the wage differential between jobs will reflect the compensating differential required for the marginal worker, not the typical worker.

Returning to the above example, note that in the equilibrium, many of the workers who are mattress testers would have to be paid a much larger wage differential than ÄW before they'd be willing to switch to that job: they are inframarginal. Thus the market compensating differential is too small for them. Conversely, the market compensating differential is more than enough for many of the workers who choose to remain toxic spill cleaners. These are the workers who have a relatively high tolerance for this kind of job. The labor market sorts workers according to their tastes for nonpecuniary aspects of the jobs. This is probably a good thing: we want the toxic spill cleaners to be the people who are least disgusted by such work. But we cannot use the compensating differential to draw conclusions about how the average or typical worker would evaluate job characteristics. This has important implications for how economists estimate the money value of life and health, which we can discuss in class.

A second set of limitations of the competitive model is that in the real world some of the assumptions underlying the model may fail to hold. For example, there simply may not be perfect mobility between occupations. Consider physicians and professional hockey players. There may be plenty of hockey players who simply could never qualify to be doctors, no matter how much time and effort they put into it. And there are surely many doctors who could never become professional hockey players. These occupations constitute *non-competing groups.*

Furthermore, *information may be imperfect*. Workers may have difficulty obtaining and evaluating information about job characteristics, work-related health and safety risks. Thus

workers may end up in jobs that are risky but do not provide adequate compensating wage differentials. The failure of perfect information provides an economic argument in favor of safety and health regulations (OSHA).

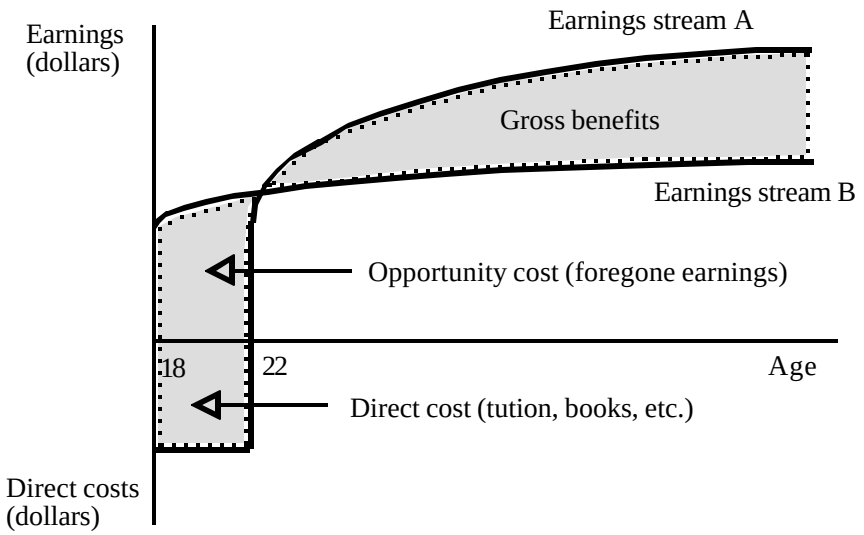
Compensating differentials applied to education and training: the human capital model

One well established feature of the labor market is that more educated workers get paid more on average. Why is this? The human capital model argues that education is a form of investment, whereby workers gain valuable skills that increase their productivity and therefore market value. Unlike investment in physical plant and equipment, education is investment in a form of capital that is intangible but critically important: namely, knowledge. Thus education can be viewed as investment in human capital.

In the basic human capital model, each individual decides how much investment in education to make, based on the balance of benefits and costs. The costs of education include both direct and indirect (opportunity) costs. The direct costs include tuition, books, and other expenses associated directly with schooling. The opportunity costs are any foregone earnings that arise from allocating time to schooling rather than working for pay.

The benefits of education come in the form of pecuniary benefits (higher earnings) and nonpecuniary benefits (being a cultured person, being able to appreciate or understand more, being more attractive to potential mates, etc.).

The nature of the tradeoff can be seen in the following diagram, which shows the hypothetical costs and benefits of a college degree over a person's lifespan. Earnings stream A represents the path of annual earnings of the person if they do not attend college. Earnings stream B represents the path of earnings for the same person if they obtain a four-year college degree. Direct costs of education are represented below the X axis.



Benefits and costs of college over the life cycle

Thinking about education as an investment gives us a simple rule for optimal investment: A person should continue to invest in one more year of education as long as the marginal benefit of one more year exceeds the overall marginal cost (MB > MC).

This rule has several implications:

- Schooling is for the young. The earlier a person obtains their education, the more years they will be able to benefit from the higher earnings. From a pecuniary point of view, it makes little sense for a person to gain new marketable skills when they are close to retirement age.
- Investment in education should be greater to the extent that the costs are lower. Thus for example government subsidies of education would make it more attractive to an individual by reducing the direct costs of attending school.
- Investment in education should be greater to the extent that the gap in earnings between more and less educated workers is greater. This makes the benefits bigger.
- Individuals should be willing to invest more in education to the extent that they are more "forward-looking." That is, a person who is willing to wait to get the returns

places greater value on the benefits of higher earnings than a very impatient individual.

These predictions are generally consistent with the evidence. Most people obtain most of their education when they are young. Subsidizing schooling tends to increase school attendance. And there is evidence that during periods when the wage gap between the more and less educated is very large (such as it is today), people choose to stay in school somewhat longer, other things equal.

Evidence on the role of impatience is more difficult to come by. Some economists suggest that a piece of evidence consistent with this prediction is the well-known inverse correlation between cigarette smoking and education. Better educated individuals might be less likely to smoke because they have learned more about the dangers of smoking. But the causation may well be the reverse. Individuals who smoke tend to be people who downplay the future costs associated with the long-run health risks, just as they undervalue the future benefits from education. Of course, it is difficult to separate out this effect from the effects of socioeconomic background and other factors that could produce the correlation.

The marginal rate of return on education

If education is an investment, is it a good one? What is the rate of return to education for an individual? A rate of return can be thought of as the increase in annual income or benefits divided by the up-front amount of capital invested. For example, if you put $100 in the bank and the bank pays you $10 per year to hold the money, the rate of return is 10/100 = 0.1 = 10%.

Turning to the case of education, suppose that the cost of one more year of schooling is D + Y, where D is direct expenses and Y is one year's income, which is the foregone earnings or opportunity cost. And suppose the benefits on an annual basis are ΔY + N, where ΔY is the boost to annual earnings provided by having one more year of schooling, and N is the annual dollar value of non-wage benefits from education.

Then the approximate marginal rate of return to one more year of schooling is the annual stream of benefits divided by the initial investment:

$$r = \frac{\Delta Y + N}{D + Y} \approx \frac{\Delta Y}{Y} \quad \textit{if } N \textit{ and } D \textit{ are relatively small}$$

ÄY/Y is the percentage change in money earnings due to one more year of schooling. We have seen that this can be estimated as the slope of an earnings equation, where the natural log of earnings is the dependent variable and years of schooling is a regressor (independent variable). The true rate of return, (ÄY+N)/(D+Y), which includes the direct costs of education as well as the nonpecuniary benefits, could be greater or smaller than the value estimated from the earnings equation.

In class we will review some of the evidence on the returns to education. Estimates of the rate of return vary between 5 and 15% per year, which makes education competitive with many other investments, as it should be if people invest in education optimally.

Because schooling is heavily subsidized, the direct costs of education to an individual are usually less than the total costs to society. Hence the individual rate of return to education may exceed the social rate of return. However, it may also be the case that education creates positive externalities that raise the social rate of return relative to the private.

Post-school investments in human capital: The return to work experience

Although there are good reasons for investing in human capital early in life, accumulation of human capital doesn't end with formal schooling. Perhaps the most important source of increases in human capital after formal schooling is learning on the job, whether through formal training or learning by doing.

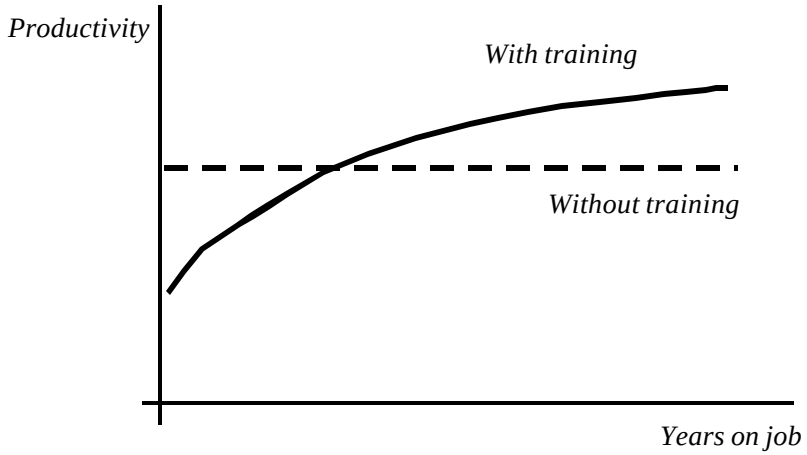
As we have seen, earnings tend to increase with age, which many economists attribute to the accumulation of human capital through work experience. Furthermore, this age-earnings profile tends to be convex. That is, earnings increase over the life cycle, but at a declining rate. The convexity is consistent with the human capital model, because most of the investment in on-the-job training, like investment in schooling, should occur early in life, so as to lengthen the period during which one receives the returns. Thus skills increase at a declining rate.

We can distinguish two kinds of training that occur on the job: general and specific.

- *General training* results in skills (*general human capital*) that can be transferred to and have value with other employers.
- *Specific training* results in skills (*specific human capital*) that are only of value to a single employer.

There are costs associated with on-the-job training, even though the individual is working,

producing, and earning as she or he learns. Generally during any learning process the worker is less productive than they would be if their entire energies and attention were focused on getting the job done. The benefits of training take the form of higher productivity later on. The following diagram shows the paths of productivity over a worker's career, with and without training on the job. Without training, the worker's productivity is constant over her career. With training, productivity increases over time (at a decreasing rate, given our prediction that more investment in training occurs early in the career).



An interesting question is who pays the costs associated with training on the job, and who reaps the benefits: the employer or the employee? If the employee paid for all the training and received all the benefits, her earnings would follow her productivity exactly, starting low (below what she would earn if she weren't training) and rising over time. If the employer paid all the costs and received all the benefits, the worker's earnings would be constant (follow the dashed line). The employer would then be paying the worker more than she was worth (in productivity terms) early in her career and less than she was worth later on. If the worker and employer shared the costs and benefits, the path of pay would lie somewhere between the two pay profiles.

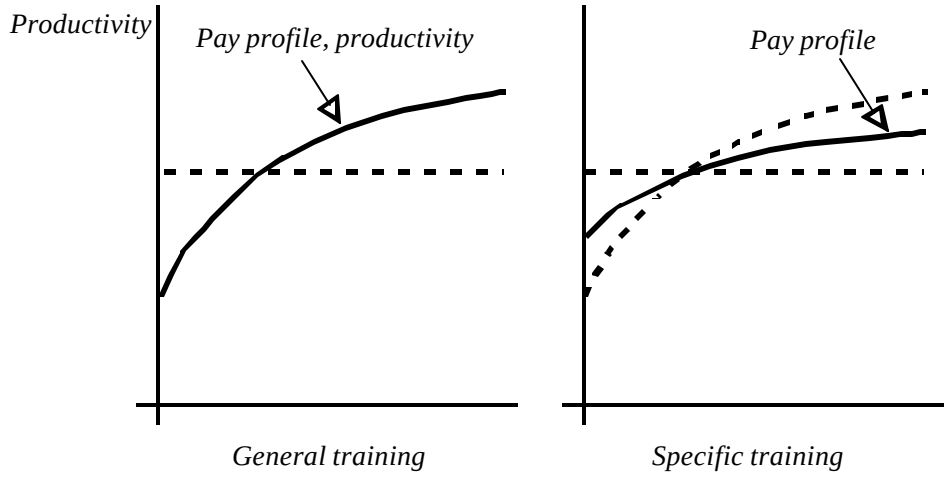
Economic theory predicts that *the worker (employee) must usually pay for any general training*. The reason is that the skills are transferrable, so the employee can retain all the benefits

of the training should she change employers. Therefore, the employer would not normally be willing to pay for that training. The fact that a worker pays all the costs and receives all the benefits of general raining implies that her pay over time should closely follow her productivity.

The prediction that employees pay for all general training appears to be contradicted by the widespread practice of employers subsidizing the formal schooling of their employees. For example, many students in Santa Clara's MBA program have their tuition paid by their employers. Since the MBA is general training (knowledge easily transferred between employers), economic theory predicts that the employee should have to pay for it. Perhaps employees actually are paying for their MBA tuition benefits in the form of lower salaries.

In contrast with the case of general training, workers and their employers share the costs and benefits of specific training. When there are specific skills, both the employee and employer have an incentive to reduce turnover, because both parties would lose their investment in specific human capital if they separated. The sharing of the costs and benefits of specific training helps accomplish this goal. To the extent that the worker is paid more because of her specific skills, she has less incentive to leave for another job, where her specific skills would have no value. But because the employer is not paying her the full value of her added productivity, the employer has an incentive to keep the worker on.

The following diagram compares the typical earnings paths for a worker receiving entirely general training and one receiving entirely specific training:



Productivity and pay over the life cycle for general and specific training

Work experience and gender differences in pay

The human capital model has interesting implications for gender differences in careers and pay. Before the 1940s or so, the vast majority of married women in the United States fit a "traditional" pattern of life-cycle work behavior of women. After marriage, most women did not work outside the home, at least not until their children were grown.

This pattern is no longer the norm, with most married women now in the labor force. Yet women generally remain the primary caregivers for children. As a consequence, married women's careers are still often interrupted for bearing and caring for children. These interruptions imply that compared with married men, married women tend to have less total work experience, less time in the workforce to reap returns to training, and career paths punctuated by periods of months or years out of the labor force.

These gender differences in work patterns are consistent with several features of women's pay and career patterns that would follow from the predictions of human capital theory.

- Women's pay is lower than men's.
- Women's earnings profiles tend to be flatter than men's, because of the diminished incentive to invest in training.

- Women appear to have pursued careers in which their skills depreciate more slowly during episodes away from the workforce. This may be one reason women have not moved as rapidly into highly technical fields, which require staying in the labor force to keep pace with rapid changes.
- As women's work experience has become more like men's, their career choices, pay, and earnings profiles have converged towards men's.

None of this is intended to deny or downplay the role of discrimination against women, such as the "glass ceiling." Rather, it is to suggest that the differences in gender roles outside the workplace is an important factor contributing to gender differences in the workplace.

The signaling model of education and earnings

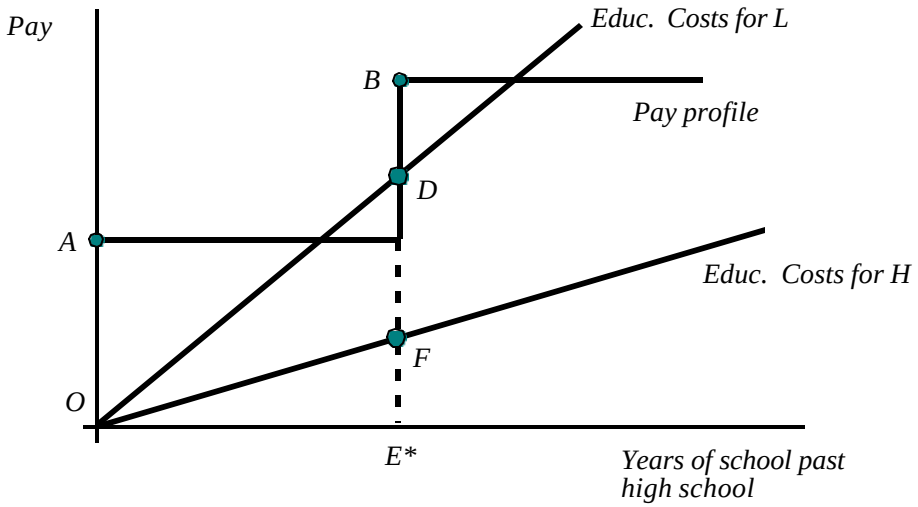
The human capital model basically argues that education pays because educated workers gain knowledge at school that is valued in the market. But there are other reasons to expect a relationship between earnings and education. An important alternative explanation is that education may provide a *signal* to employers of an individual's productivity, when that productivity is difficult to observe. In this view, *an educated person earns more not because of what they have learned but because of what their schooling tells employers about them.* If signaling were the only effect of education, it would imply that education does not increase productivity but rather helps identify the types of individuals who are likely to have greater productivity regardless of what they learned in school.

What does schooling signal that employers might value? A person who has more education may be more likely have greater cognitive ability, drive, perseverence, and diligence. They may also be more disposed to submit to authority. In this regard signaling is different from the idea that education is a pure credential that provides social contacts or status in an "old boy network."

Of course, education could have *both* signaling and human capital effects on earnings (and probably does have both).

The signal is only valuable if, on average, the workers who choose to get more education ("self-select") are also the ones who are more productive. This requires that productivity and the costs of education tend to be *negatively correlated.* If this weren't true, then everyone would choose to buy the signal, and then it would be of no value to employers.

The following diagram illustrates a simple signaling equilibrium. The horizontal axis represents the number of years of education past high school. Two kinds of workers are depicted: a low-productivity worker (L) and a high-productivity worker (H). The sloped lines represent the total costs of education (on an annualized basis) for each worker type. Schooling beyond high school is more costly to the low-productivity worker (presumably because it requires much greater effort and pain to get through college).



Signaling model of educational choice

The pay profile shows how pay varies by education. In this simple example, we assume that employers offer to pay a lower wage (at A) to workers with less than E* years of education beyond high school, and up to B for workers with E* years or more. The employers cannot directly observe whether a worker is type L or type H, but they can hope that workers will reveal which type they are through their choice of years of education.

To decide whether to obtain additional schooling beyond high school, the low-skilled worker compares her pay without the additional schooling against her pay net of college costs if she obtains E* years of college. The distance OA is her pay with just the high-school diploma; the distance DB is the benefit of the higher pay with E* years of college, net of her costs of education. Because OA > DB, this worker prefers not to invest in any additional schooling.

The high-productivity worker H makes a similar comparison, but for her the net gain due to E* years of college is FB. Since FB > OA, this additional schooling is worth it to the high-productivity worker.

Thus if employers have set this sort of pay profile, high-productivity workers obtain E* years of additional schooling, low-productivity workers would not, and employers would be able to use education as a signal of the underlying ability (productivity) of a worker. In class we will note that the same result could be obtained with slight changes in the location E* or the shape of the pay profile. We will also see that signaling would not work if the positions of the curves representing schooling were reversed for workers L and H (positive correlation between educational costs and productivity).

Both the signaling and the human capital models predict a positive correlation between education and earnings for an individual. Therefore it is not possible to decide between the two models based on that observed correlation in the data. There is little reason to doubt that both factors are at work in the real world. But there is considerable dispute about their relative importance. In class we will discuss some of the evidence economists use to try to quantify their relative importance.

In the signaling model, education need not contribute to the productivity of the worker, so one might question whether education would have any value to society at all. In fact it would, because under signaling education acts as a sorting device, helping employers match workers' skills with the requirements of the job. But the signaling model also implies that once the sorting has occurred, further investments in education may have little *economic* payoff. This is a further reason that the individual and social rates of return to schooling might differ.

Institutional models of differences in earnings: an overview

There are many other factors affecting pay differences across individuals besides schooling and work experience. You can see from your earnings regressions that education and experience yield an R-squared of about 20%, leaving 80% of the variation in earnings across individuals unexplained by these factors. Some important additional factors contributing to variation in earnings are institutional differences across jobs and labor markets. By "institutional" differences I mean differences in the way markets behave that are due to their rules and organizational structure rather than pure supply and demand. Among these are:

Unions and bargaining. In the United States, union membership as a percentage of nonagricultural employment has fallen off from its peak in the 1950s of 33-34% to about 15% now, though unions are still quite important in some parts of the economy, such as the public sector and transportation. Most research finds that unions do have a positive effect on the wages of their members, though they may also depress the wages of non-members.

Interindustry wage differentials. Data show that there are persistent wage differences between industries, even after controlling for individual education and experience. There are various possible explanations for these patterns, some of which we will discuss in class.

Discrimination. Some employers discriminate on the basis of gender, race, ethnicity, age, etc. The effect of discrimination on wages for these groups depends on many factors, including whether or not there are enough jobs available with employers who do not discriminate.