1 A Very Brief Math Review

All the math you need for this course is summarized here.

1.1 Sigma Notation

\[ \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + \cdots + Y_{N-1} + Y_N \]

\( \sum \) is called sigma notation. \( \Sigma \) is the Greek capital letter sigma (\( \sigma \) is the lower case sigma.). This symbol indicates a summation of a series of variables. In this case, the notation says to start summing the \( Y_i \) variables starting with \( Y_{i=1} \) and ending at \( Y_{i=N} \).

1.2 Exponents

For any \( X, a, b, c \),

i)

\[ X^{-a} = \frac{1}{X^a} \]

and

\[ \frac{1}{X^{-b}} = X^b \]

ii)
\[ X^a X^b = X^{a+b} \]

iii)

\[ (X^a)^b = X^{ab} \]

iv)

\[ (X^a Y^b)^c = X^{ac} Y^{bc} \]

v)

\[ X^0 = 1 \]

\[ X^1 = X \]

\[ X^{-1} = \frac{1}{X} \]

\[ X^{1/2} = \sqrt{X} \]

\[ X^{-1/2} = \frac{1}{\sqrt{X}} \]

\[ X^{-1/3} = \frac{1}{\sqrt[3]{X}} \text{ (cube root)} \]
1.3 Logarithms

The natural logarithm is given by \(\ln\) or sometimes \(\log\).

*Rules for Logarithms*

For any \(X\) and \(Y\):

i) \(\ln XY = \ln X + \ln Y\).

ii) \(\ln \frac{X}{Y} = \ln X - \ln Y\).

iii) \(\ln X^Y = Y \ln X\).

iv) \(\ln 1 = 0\).

v) \(\ln 0\) undefined.

1.4 Calculus

The partial derivative of a function \(f(X, Y)\) with respect to \(X\) is

\[
\frac{\partial f(X, Y)}{\partial X}.
\]

If there is only one argument of a function, the derivative is sometimes given by this notation:

\[
\frac{\partial f(X)}{\partial X} = f'(X).
\]
Rules for partial derivatives:

i) if $f(X, Y) = d + aX$, where $a$ and $d$ are constants, then

\[
\frac{\partial f(X, Y)}{\partial X} = a.
\]

Example:

\[
\frac{\partial (3X)}{\partial X} = 3
\]

ii) if $f(X, Y) = d + aX^b$, where $a$, $b$, $d$ are constants, then

\[
\frac{\partial f(X, Y)}{\partial X} = abX^{b-1}.
\]

Example:

\[
\frac{\partial (2 + 3X^5)}{\partial X} = 3 \times 5 \times X^4 = 15 \times X^4
\]

iii) if $f(X, Y) = d + aX^bY^c$, where $a$, $b$, $c$, and $d$ are constants, then

\[
\frac{\partial f(X, Y)}{\partial X} = abX^{b-1}Y^c.
\]

Example:
\[
\frac{\partial (2 + 3X^5Y^7)}{\partial X} = 3 \times 5 \times X^4Y^7 = 15X^4Y^7
\]

iv) if \( f(X, Y) = d + aX^bY^c \), where \( a, b, c \), and \( d \) are constants, then

\[
\frac{\partial f(X, Y)}{\partial Y} = aX^bY^{c-1}.
\]

Example:

\[
\frac{\partial (2 + 3X^5Y^7)}{\partial Y} = 3 \times 7 \times X^5Y^6 = 21X^5Y^6
\]

v) if \( f(X, Y) = \ln X \),

\[
\frac{\partial f(X, Y)}{\partial X} = \frac{1}{X}.
\]

vi) Chain Rule: if \( f(X, Y) = g(h(X)) \), then

\[
\frac{\partial f(X, Y)}{\partial X} = \frac{\partial g(h(X))}{\partial h(X)} \frac{\partial h(X)}{\partial X}.
\]

An example: if \( f(X, Y) = \ln(aX^b) \), where \( a \) and \( b \) are constants, then

\[
\frac{\partial f(X, Y)}{\partial X} = \frac{1}{aX^b} abX^{b-1} = \frac{b}{X}.
\]
Note: \( g(\cdot) = ln(aX^b) \) and \( h(X) = aX^b \).

**What Do they Mean?**

Derivatives indicate the relationship between a function and a variable. Consider the case of a line

\[
Y = f(X) = a + bX.
\]

The derivative of \( Y \), which is a function of \( X \), is simply the slope of this line.

\[
\frac{\partial f(X)}{\partial X} = b
\]

If the slope is positive \((b > 0)\), the derivative is positive. \( X \) increases \( Y \).

If the slope is negative \((b < 0)\), the derivative is negative. \( X \) decreases \( Y \).

If the slope is zero \((b = 0)\), there is no relationship between \( X \) and \( Y \). This is a flat line.

**An Example**

Consider a line:

\[
Y = f(X) = 4 + \frac{1}{2} X.
\]

If \( X \) increases by 6, how is \( Y \) affected?

The derivative of \( Y \) with respect to \( X \) is
\[
\frac{\partial f(X)}{\partial X} = \frac{1}{2}
\]

If \( X \) increases by 6, \( Y \) changes by

\[
\frac{\partial f(X)}{\partial X} \cdot \Delta X = \frac{1}{2} \cdot 6 = 3.
\]

1.5 Optimization

An unconstrained optimization problem:

\[
\max_{X,Y} f(X, Y),
\]

where \( X \) and \( Y \) are the two choice variables, and \( f(X, Y) \) is the objective function.

Solving an Optimization Problem

To solve this problem, derive the first order conditions for maximization.

1) With respect to \( X \)

\[
\frac{\partial f(X, Y)}{\partial X} = 0
\]

2) With respect to \( Y \)
This provides a system of two equations and two unknown variables (\(X\) and \(Y\)). This system can be solved for the optimal values of \(X\) and \(Y\), called \(X^*\) and \(Y^*\).

A constrained optimization problem:

\[
\max_{X,Y} f(X,Y) \quad s.t. \quad g(X,Y) = C,
\]

where \(g(X,Y) = C\) is the constraint in the problem.

In this course, the constraints are simple enough where we can find a way to re-write the problem as an unconstrained problem. This is generally accomplished by substituting the constraint into the objective function.
2 Overview of the Labor Market

2.1 Labor Markets

A labor market is where some unit of labor is bought and sold. A unit of labor may be an hour, year, or some other measure. Labor is a good with a price called the wage.

There are many different kinds of labor (labor of college educated workers, labor of high school educated workers, labor of experienced workers, labor of inexperienced workers, labor of medical doctors, labor of waiters, etc.). For each kind of labor there is a market. A market may have geographic dimensions. The labor market for doctors in India may be a separate labor market from the labor market for doctors in the United States.

2.2 Labor Market Participants

Workers

Workers supply labor. Workers are the employees of firms. Workers make decisions about whether to work and how much to work. Workers receive wages for their work, which they use to purchase consumer goods.

Firms

Firms demand labor. Firms are the employers of workers. Firms combine labor with other inputs, such as capital (machines, factories, etc.), to produce
output goods. Firms can be one person (the worker is self-employed) or very large with thousands of employees.

**Government**

Various levels of government (local, state, federal) enforce laws which regulate wages (e.g. minimum wage laws), who can work (e.g. children cannot work in the United States), and working conditions (e.g. construction workers must wear hard hats). The government also taxes wages and subsidizes individuals who do not or cannot work (e.g. Social Security benefits for disabled workers).

**Unions**

Unions are groups of workers who collectively bargain labor contracts with employers. Usually unions are formed among workers in the same industry or occupation, for example the United Auto Workers (UAW) or the American Federation of Teachers (AFT).

### 2.3 Measurement

Measuring employment, hours worked, and wages is far from straightforward. There are many potential definitions. Some definitions are better suited for some populations than others. In empirical applications, measurement is dictated by data limitations, as surveys may only collect data according to one set of measurements.
2.3.1 Where Does Data Come From?

Data used in labor economics typically comes from three sources: 1) surveys of individual who self-report information about how much they work and how much they get paid, 2) surveys of firms in which the employers report similar information for their employees, 3) government statistics (e.g. tax returns or social security claims). Federal, state, and local governments spend millions of dollars each year to fund labor force surveys. The United States decennial Census is probably the most well known survey of individuals. Every 10 years, the Census Bureau surveys all household in the United States, although only a subset fill out the detailed “long form” of the Census.

The main source of labor force statistics in the United States is a survey called the Current Population Survey (CPS) conducted by the Bureau of Labor Statistics (BLS). Each month the survey contacts about 60,000 households (usually by phone) and asks them a series of questions about their labor market activities in a recent week. From these questions, the BLS constructs a number of statistics, the most widely cited of which is the unemployment rate.

Measurement Errors

It is important to keep in mind that much of the data used in labor economics is prone to several kinds of errors. Because the data is self-reported by individuals voluntarily, some people will not respond. Having at least 20-30 percent of individuals refuse to respond to a survey is not uncommon. An-
other source of error occurs because the survey asks about past employment, which respondents may not remember correctly. Do you know exactly how many hours you worked last week or exactly how much you were paid last year? Finally, some respondents have an incentive to intentionally misreport information, even though most surveys like the Census and CPS have strict confidentiality rules. For example, a respondent may want to under-report her wages because of the fear that this information would be reported to the Internal Revenue Service (it’s not).

2.3.2 Employment

The BLS divides the United States population into three mutually exclusive groups: employed people, unemployed people, and people defined as “out of the labor force” (OLF). Only adults 16 or older are counted; children are excluded from all three groups.

Employed An individual is considered employed if they worked at a job for at least 1 hour (self-employed are included) or worked at least 15 hours on a nonpaid job (e.g. family farm). This latter part of the definition is difficult to check, but generally this excludes people working at home for no pay (e.g. mothers or fathers taking care of their children).

Unemployed An individual is considered unemployed if they are temporarily unemployed (e.g. construction workers not working because of bad weather) or have been actively looking for work in the past 4 weeks. Again, this
definition is somewhat vague and arbitrary. How do we define “actively” looking for work?

*Out of the Labor Force* Anyone not considered either employed or unemployed is counted as out of the labor force. This generally includes students without jobs, homemakers, retirees, etc.

**Definitions**

\[
\begin{align*}
\text{Population} &= \text{employed} + \text{unemployed} + \text{out of labor force} \\
\text{Labor Force} &= \text{employed} + \text{unemployed} \\
\text{Labor Force Participation Rate} &= \frac{\text{labor force}}{\text{population}} \\
\text{Unemployment Rate} &= \frac{\text{unemployed}}{\text{labor force}}.
\end{align*}
\]

The labor force participation rate in the United States is about 75 percent for all men and 60 percent for all women. The labor force participation rate varies substantially with age. For men age 25-44, the labor force participation rate is about 90 percent.

The current unemployment rate is about 4-5 percent in the United States. In many countries in Western Europe, the unemployment rate is around 10 percent.

*Why are these statistics sometimes misleading?*

Notice that the unemployment rate is affected by the size of the labor force. However, individuals may enter or exit the labor force at any time. During economic recessions, unemployed workers may stop “actively” looking for work and be counted as out of the labor force. This *discouraged*
worker effect is used to explain why the measured unemployment rate understates the real economic conditions. Similarly, as the economy recovers from a recession, workers out of the labor force may start looking for work. This would factor would tend to overstate the measured unemployment rate during an economic recovery.

2.3.3 Hours

Ideally labor economists would want to measure all labor by a continuous variable, such as the number of hours worked. For workers who are paid by the hour, this information is readily available. For workers paid a salary or paid by the task (e.g. salesman paid sales commissions), the number of hours worked is not so easily measured. I am paid a salary and I would have difficulty reporting on a survey the exact number of hours I worked last week.

A related issue is that a continuous measure of hours may not be the appropriate measure of labor supply. Few workers actually adjust their hours continuously in response to changes in wage rates. If an employer offers a lower wage, workers may simply quit the job and move from 40 hours to 0 hours. Many economic models would assume a worker smoothly adjusts her hours by moving from 40 hours to 34 hours, for example. Later in the course, we will discuss models that incorporate this idea of discontinuous labor supply.

Firms, too, have a related problem. Firms generally hire workers not hours of labor. Because of fixed hiring and firing costs, firms may also have
difficulty smoothly adjusting their labor demand.

2.3.4 Wages

One of the most challenging measurement issues in labor economics is measuring an individual’s pay. The ideal situation would be if everyone was paid a constant wage rate for each hour. Unfortunately, this is only the case for some hourly workers. Calculating a worker’s hourly wage is difficult if the worker is paid a salary, a sales commission, or receives tips.

Total Compensation

Wages are only one element of the total compensation workers receive. Workers often receive a large share of their compensation in the form of benefits, e.g., health insurance and pensions. Accurate information on benefits is often difficult to obtain from respondents to surveys. And many surveys do not ask for this information (e.g., there is only limited information in the CPS on benefits).

Workers may also receive utility from simply working at the job. An employer’s investment in the working conditions (e.g., new office furniture) can change the level of utility a worker receives from a job. For many jobs, an investment in protecting worker safety (e.g., installing smoke detectors) can also increase the utility from the job by reducing the risk of injury or death.

To the extent that the benefits and other aspects of utility from working at
a job are not counted, wages may understate or overstate the utility workers receive from a job. If the level of these benefits varies systematically across occupations and industries (e.g. few benefits for restaurant workers, many benefits for accountants), wages may fail to capture the true variation in the returns that workers receive from various jobs.

*Real vs. Nominal Wages*

An issue that often arises in tracking trends in wages over time is measuring the purchasing power of wages as the prices of goods and services people purchase changes. *Real wages* are wages adjusted for price changes. Since prices have generally been increasing over time in the recent history, real wages are adjusted for inflation. *Nominal wages* are the actual wage paid in any given period of time.

We generally use a price index based on a representative bundle of goods to measure price changes. For the United States, this is most often one of the price index series which is part of the Consumer Price Index (CPI). Price indexes work by picking a base year, say 2005, and expressing all other prices in terms of this base year.

An example: Normalize our base year price to $p_{2005} = 100$. If there was 3 percent inflation from 2005 to 2006, our price index increased from $p_{2005} = 100$ to $p_{2006} = 103$. We can construct a price deflator as

$$
\frac{p_{2005}}{p_{2006}} = \frac{100}{103} = 0.971
$$
We can deflate 2006 nominal wages into 2005 dollars by multiplying the 2006 nominal wages by this price deflator. A worker is paid a nominal wage of $w_{2006} = 11$ in 2006 and a nominal wage of $w_{2005} = 10$ in 2005. In nominal terms, the worker’s wage increased by 10 percent.

The real wage in 2006 (in 2005 dollars) is

\[
\frac{w_{2006}^R}{w_{2006}} = \frac{w_{2006}}{p_{2005} \cdot p_{2006}} = 11 \cdot \frac{100}{103} = 10.68
\]

Given inflation, real wages only increased 6.8 percent from 2005 to 2006, although nominal wages increased 10 percent. We would argue that the real wage increase better represents what how much this worker’s living standards increased, as the real wage indicates how many more goods and services she can purchase.

### 2.4 Current Statistics for the United States

Take a look at the wealth of information on the BLS webpage.

http://www.bls.gov

### 2.5 Supply, Demand, and Equilibrium in the Labor Market

We can readily apply the models learned in introductory microeconomics to the labor market. As we will discuss later, labor is a unique good and
the supply and demand for labor will receive specialized treatment in future lectures. For now, we can think of an hour of labor as any other good, like a bushel of wheat.

Label $h$ as the aggregate hours of labor in the market. The price of labor is the wage rate $w$.

Figure 1 displays a standard supply and demand graph with $w$ replacing price $p$ on the vertical axis and $h$ replacing quantity $q$ on horizontal axis. As in the market for wheat, the demand curve for labor is downward sloping. The higher the price of labor, the less firms demand of it. The demand curve is the sum of the labor demands of many individual firms. The supply curve is upward sloping. The higher the price of labor, the more workers will supply of it. The supply curve is the sum of the labor supply of many individual workers.

The intersection of the labor supply and labor demand curves provides the equilibrium wage ($w^*$) and labor employed ($h^*$). All workers are paid $w^*$ and the total number of hours worked in the economy is $h^*$.
3 Labor Demand

3.1 Model of Firm Demand for Labor

Firms purchase capital and labor in the input markets. (Note: input markets are sometimes called “factor markets” for “factors of production”). Firms combine these inputs to make output goods, which it sells in the output market.

For simplicity, we assume there is only one kind of labor measured in labor hours $h$. Labor is paid only one wage $w$ for each hour. Later, we will relax these assumptions and examine a model with different kinds of labor (e.g. labor of experienced and unexperienced workers) and different wage rates paid to each kind of labor.

It is important to note that the firm is involved in three separate markets:

i) Labor Market: the firm is a consumer in this market.

ii) Capital Market: the firm is a consumer in this market.

iii) Output or Product Market: the firm is a supplier in this market.

We first examine a model in which each market is competitive: competitive labor market, competitive capital market, and competitive output market. Firms cannot influence input or output prices, and take these prices as given or constant (they are “price takers”). Later, we will examine non-competitive markets (e.g. monopsony in the labor market).
Competitive Markets Assumptions:
1) Firms can hire unlimited labor hours, $h$, at a constant wage rate $w$.
2) Firms can rent unlimited amounts of capital, $k$, at a constant rental rate $r$.
3) Firms can produce one good, $q$, and sell unlimited amounts of this good at a constant price $p$.

Are Firm Profits Zero? General vs. Partial Equilibrium

Although it does not play a large role in the discussion that follows, it should be noted that firms can earn positive profits in this model. This is a partial equilibrium model and focuses on how changes in given input and output prices affect a firm’s demand for labor and capital. In the partial equilibrium model, input and output prices are exogenous or given. In a general equilibrium model, the input and output prices may respond to the labor and capital demand decisions of the firms in the economy. In a general equilibrium model, input and output prices are endogenous, and reflect the level of market demand for inputs and the level of market output. In a general equilibrium model, it is possible that competition among firms would lower profits to near zero for individual firms.

Production Function

Output is produced according to this production function. The production function embodies the technology of production.
\[ q = f(h, k) \]

Output is increasing in labor hours and capital.

\[ \frac{\partial f(h, k)}{\partial h} > 0 \]

\[ \frac{\partial f(h, k)}{\partial k} > 0 \]

**Labor Demand of a Firm and Market Demand**

For the moment, we consider a representative firm. However, the market demand for labor comes from a number of potentially heterogenous firms. In general, firms can differ in their production technologies (production functions), the output goods they produce, and the prices they receive for their output goods. Firms that sell hot dogs and firms that manufacture pencils both demand labor, but differ in many dimensions. The labor demand curve for a kind of labor is the aggregation of the labor demand of these many different firms.

### 3.2 Profit Maximization

Profits, \( \pi \), are defined as total revenue minus total costs:
\[ \pi = pq - wh - rk. \]

Firms choose \( h \) and \( k \) to maximize profits. Firms do not choose prices. Firms also do not choose output. Output is a function of input choices as given by the production function.

The firm’s maximization problem is

\[
\max_{h,k} \quad pq - wh - rk \quad \text{s.t.} \quad q = f(h, k),
\]

where \( \text{s.t.} \) is “subject to” and indicates that the production function is the constraint in this problem.

The simplest way to solve this problem is to substitute the production function for \( q \). This yields an unconstrained maximization problem:

\[
\max_{h,k} \quad pf(h, k) - wh - rk.
\]

Now derive the first order conditions.

1) Write the partial derivative with respect to \( h \):

\[
p \frac{\partial f(h, k)}{\partial h} = w.
\]

2) Write the partial derivative with respect to \( k \):
\[ p \frac{\partial f(h, k)}{\partial k} = r. \]

With \( w, r, p \), and an assumed form for the production function, we can solve these two equations for the two unknown variables \( h^*(w, r, p) \) and \( k^*(w, r, p) \). \( h^*(w, r, p) \) and \( k^*(w, r, p) \) are the optimal amounts of labor hours and capital units the firm would purchase. These optimal values are functions of the input and output prices. The firm’s demand for labor and capital changes as these prices vary.

It should also be noted that from the optimal labor and capital demand, the profit maximizing output can be calculated by substituting these values into the production function.

\[ q^* = f(h^*(w, r, p), k^*(w, r, p)) \]

The profit maximizing levels of total revenue \( (TR = pq^*) \), total costs \( (TC = wh^*(w, r, p) + rk^*(w, r, p)) \), and profits can all be calculated by substituting the optimal levels of labor and capital.

### 3.3 Some Terminology

*Marginal Product of Labor*
\[ MP_h = \frac{\partial f(h, k)}{\partial h} \]

\textit{Marginal Product of Capital}

\[ MP_k = \frac{\partial f(h, k)}{\partial k} \]

\textit{Marginal Rate of Technical Substitution}

\[ MRTS = \frac{MP_h}{MP_k} \]

\textit{Marginal Revenue Product of Labor}

\[ MRP_h = pMP_h \]

\textit{Marginal Revenue Product of Capital}

\[ MRP_k = pMP_k \]
3.4 What Do the First Order Conditions Mean?

Given the definitions of marginal product of labor and capital, we can rewrite the first order conditions as

1)\[ pMP_h = w, \]

or\[ MRP_h = w. \]

2)\[ pMP_k = r, \]

or\[ MRP_k = r. \]

\( MRP_h \) provides the value in additional revenue of one more labor hour. Each additional labor hour produces \( MP_h \) * 1 of output. This output can be sold in the output market at price \( p \). The wage rate provides the cost of one more labor hour. The marginal cost of labor is \( w \).

The intuition behind the first order conditions is that the firm should
continue to purchase labor and capital inputs up until the point that the marginal benefit of these inputs ($MRP_h$ and $MRP_k$) equals the marginal cost, $w$ and $r$. The first order conditions are simply a re-statement of a fundamental principle of economics: decision makers make choices by equating the marginal benefit with the marginal cost.

At the point where $MRP_h > w$, the benefit of more labor hours exceeds the cost. The firm should hire more labor and increase profits.

At the point where $MRP_h < w$, the cost of more labor hours exceeds the benefit. The firm is losing money at this point and should hire less labor.

### 3.5 Labor Demand Elasticity (with respect to wages)

The labor demand elasticity with respect to wages is defined as

$$
\epsilon = \frac{\% \Delta h^*(w, r, p)}{\% \Delta w}.
$$

or

$$
\epsilon = \frac{\partial h^*(w, r, p)}{\partial w} \frac{w}{h^*}.
$$

(To see the connection between these two definitions of elasticity, note that the percent change in $X$ from $X_0$ to $X_1$ is $\frac{X_1 - X_0}{X_0} 100\% = \frac{\Delta X}{X_0} 100\%$.)

Another way to write this is using natural logs
\[ \epsilon = \frac{\partial \ln h^*(w, r, p)}{\partial \ln w} \]

Given the assumptions of profit maximization, \( \epsilon \leq 0 \). This is equivalent to the assumption that the labor demand curve is downward sloping.

The labor demand elasticity has the following interpretation:

i) \( \epsilon = 0 \) (perfectly inelastic labor demand). If wages increase by \( X \% \), labor demand does not change.

ii) \( -1 < \epsilon < 0 \) (inelastic labor demand). If wages increase by \( X \% \), labor demand decreases by less than \( X \% \).

iii) \( \epsilon = -1 \) (unit elastic labor demand). If wages increase by \( X \% \), labor demand decreases by exactly \( X \% \).

iv) \( \epsilon < -1 \) (elastic labor demand). If wages increase by \( X \% \), labor demand decreases by more than \( X \% \).

v) \( \epsilon = -\infty \) (perfectly elastic labor demand). If wages increase by \( X \% \), labor demand goes to 0 after the wage increase.

It is important to note that \( \epsilon \) is (in general) a function of input prices (\( w \) and \( r \)), level of labor hours (\( h \)) and units of capital (\( k \)) already used, and the output price (\( p \)). This means that the labor demand elasticity can vary as these factors change.
3.6 Scale and Substitution Effects

We can decompose how changes in wages affect labor demand into two factors. Consider a reduction in the wage rate (and everything else remains the same, including capital and output prices).

1) Scale Effect

Lowered wage rates decreases the cost of production. This induces firms to increase output and demand more labor. This effect is called a scale effect and is analogous to an income effect in consumer theory (e.g. price of apples declines, the consumer is in effect “wealthier”, so she purchases more of all goods).

2) Substitution Effect

Lowered wages relative to capital rental rates makes labor relatively less expensive compared to capital. This induces the firm to shift its input mix toward labor, and away from capital, and therefore demand more labor. The substitution effect is analogous to the substitution effect in consumer theory as the relative prices of consumer goods change (e.g. price of apples declines relative to oranges, so consumers purchase more apples).

3.7 Elasticity of Substitution

The elasticity of substitution indicates how easily firms can change their input mix as relative input prices change. The elasticity of substitution
is determined by the production function. The elasticity of substitution is defined holding output constant (no scale effects):

\[ \sigma = \frac{\% \Delta (k/h)}{\% \Delta (w/r)}, \]

or

\[ \sigma = \frac{\partial (k/h)}{\partial (w/r)} \frac{w/r}{k/h}. \]

Or in logs,

\[ \sigma = \frac{\partial \ln(k/h)}{\partial \ln(w/r)}. \]

Notice that for profit maximization, \( w/r = MRTS \). Recall the first order conditions for profit maximization:

\[ MRTS = \frac{MP_h}{MP_k} = \frac{w}{r}. \]

Substituting this into the definition of \( \sigma \), the elasticity of substitution can also be written as

\[ \sigma = \frac{\% \Delta (k/h)}{\% \Delta (MRTS)}, \]

or
\[ \sigma = \frac{\partial (k/h)}{\partial (MRTS)} \frac{MRTS}{k/h}. \]

Or in logs,
\[ \sigma = \frac{\partial \ln(k/h)}{\partial \ln(MRTS)}. \]

\( \sigma \geq 0 \). Higher values of \( \sigma \) indicate that the firm can more easily substitute inputs as the relative input prices change. For example, a firm with a high elasticity of substitution would respond to an increase in \( \frac{w}{r} \) by rapidly substituting capital for labor (e.g. replacing workers with machines).

It should be clear then that a firm with a production technology characterized by a high elasticity of substitution should also have a high labor demand elasticity with respect to wages (high \( \sigma \) implies a high \( \epsilon \)). A firm that can easily substitute between labor and capital would respond to a change in wage rates by rapidly shifting between labor and capital and rapidly changing its labor demand.

### 3.8 Types of Production Functions

Firms have different types of production technologies indicated by different production functions. There is likely a considerable amount of variation in the technology of production among firms in different industries (e.g. man-
ufacturing vs. services). Even facing the same input prices, a firm manufacturing cars uses a very different combination of labor and capital than a firm that produces accounting services.

3.8.1 Isoquant Curves

Figures 2a, 2b, 2c graph different types of production functions. In these graphs, capital \((k)\) is measured along the vertical axis and labor \((h)\) along the horizontal axis. The curve in each of the figures is an isoquant curve. Along this curve production of the output good is constant at \(q_0\). The equation for an isoquant curve is

\[
q_0 = f(h, k)
\]

The shape of the isoquant curve depends on the functional form of the production function. The slope of this curve is \(-MRTS\). The two extreme cases are perfect substitutes and perfect complements.

3.8.2 Perfect Substitutes

Figure 2a graphs a perfect substitutes isoquant:

\[
q_0 = f(h, k) = ah + bk,
\]

where \(a \geq 0\) and \(b \geq 0\) are constants.
An example of a perfect substitutes isoquant:

\[ q_0 = f(h, k) = 2h + 3k, \]

A perfect substitutes isoquant is a straight line with equation

\[ k = \frac{q_0}{b} - \frac{a}{b} h. \]

The vertical intercept of this line is \( q_0 / b \). The horizontal intercept is \( q_0 / a \).

In the perfect substitutes case, labor and capital inputs can be substituted by firms at a constant rate. Notice that the slope of this line \(-a/b\) is \(-MRTS\).

The elasticity of substitution for a perfect substitutes production function is \( \sigma = \infty \).

### 3.8.3 Perfect Complements

Figure 2b graphs a perfect complements isoquant:

\[ q_0 = f(h, k) = \min(ch, dk), \]

where \( c \geq 0 \) and \( d \geq 0 \) are constants.

An example of perfect complements:

\[ q_0 = f(h, k) = \min(\frac{1}{2} h, 4k). \]
Production is given by *fixed proportions* of labor and capital. In the perfect complements case, production is limited by the smallest input:

If $dk < ch$, then $q_0 = dk$. In this case, adding additional labor does not increase production.

If $dk > ch$, then $q_0 = ch$. In this case, adding additional capital does not increase production.

A perfect complements isoquant is an “L” shaped curve with a vertical line at $h = q_0/c$ and a horizontal line at $k = q_0/d$. The slope at the vertex of the isoquant is $-c/d$, which is $-MRTS$.

The elasticity of substitution for a perfect complements production function is $\sigma = 0$.

### 3.8.4 Cobb-Douglas

Another type of production function is called a Cobb-Douglas production function:

$$q_0 = f(h, k) = h^{\theta_1} k^{\theta_2},$$

where $\theta_1 \geq 0$ and $\theta_2 \geq 0$ are constants.

Figure 2c graphs a Cobb-Douglas isoquant.

The elasticity of substitution for a Cobb-Douglas production function is $\sigma = 1$. 
An example of a Cobb-Douglas isoquant is

\[ q_0 = f(h, k) = h^{1/2}k^{1/4}. \]

This production function is a middle ground production between perfect substitutes and perfect complements. The shape of the production function depends on the values of \( \theta_1 \) and \( \theta_2 \). The slope of the curve is \(-MRTS\). Derive \( MRTS \) as

\[ MRTS = \frac{MP_h}{MP_k} = \frac{\theta_1 h^{(\theta_1 - 1)}k^{\theta_2}}{h^{\theta_1}(\theta_2 k^{(\theta_2 - 1)} = \frac{\theta_1 k}{\theta_2 h}. \]

The slope of the Cobb-Douglas isoquant is then

\[ -\frac{\theta_1 k}{\theta_2 h}. \]

Unlike the perfect substitutes or perfect complements isoquants, the slope of the Cobb-Douglas isoquant changes depending on the values of \( k \) and \( h \). The slope increases (in absolute value) as the ratio of capital to labor \((k/h)\) increases.

### 3.8.5 Calculating the Elasticity of Substitution

Can we show that the elasticity of substitution for the Cobb-Douglas production function is \( \sigma = 1? \)
Use the log form of $\sigma$ with $MRTS$ in the definition:

$$\sigma = \frac{\partial \ln(k/h)}{\partial \ln(MRTS)}.$$

What is $\ln MRTS$?

From above the $MRTS$ for the Cobb-Douglas production function is

$$MRTS = \frac{\theta_1 k}{\theta_2 h}.$$

Taking logs,

$$\ln MRTS = \ln\left(\frac{\theta_1}{\theta_2}\right) + \ln\left(\frac{k}{h}\right).$$

Re-arranging,

$$\ln\left(\frac{k}{h}\right) = \ln MRTS - \ln\left(\frac{\theta_1}{\theta_2}\right).$$

Therefore,

$$\sigma = \frac{\partial \ln(k/h)}{\partial \ln(MRTS)} = 1.$$

Note that $\sigma$ is always 1 for the Cobb-Douglas production function. $\sigma$ in this case does not depend on the parameters $\theta_1$ or $\theta_2$. 
3.9 A Cobb-Douglas Example

An example: \( q = f(h, k) = h^{1/4}k^{1/4} \).

Marginal product of labor:

\[
MP_h = \frac{1}{4}h^{-3/4}k^{1/4}.
\]

Marginal product of capital:

\[
MP_k = \frac{1}{4}h^{1/4}k^{-3/4}.
\]

Profit maximizing combination of labor and capital:

First order conditions:

1)

\[
\frac{1}{4}h^{-3/4}k^{1/4} = w.
\]

and

2)

\[
\frac{1}{4}h^{1/4}k^{-3/4} = r.
\]

Divide 1) and 2):
\[
\frac{k}{h} = \frac{w}{r}.
\]

Solve for \( k \):

\[
k = h \frac{w}{r}
\]

Now substitute this equation back into either first order condition.
Substitute into first order condition 1):

\[
\frac{1}{p} \frac{h^{-3/4}}{(h^{w/r})^{1/4}} = w.
\]

Solve for \( h \) and simplify:

\[
\frac{1}{p} \frac{h^{-3/4} (hw/r)^{1/4}}{h} = w
\]

\[
h^{-3} h^{1+\frac{w}{r}} = w^4 p^{-4} r^4
\]

\[
h^{-2} = w^3 p^{-4} r^4
\]
Finally,

\[ h^* = \frac{1}{16} \frac{p^2}{r^{1/2}w^{3/2}} \]

Labor demand is a function of output prices, price of capital, and wage rates. Check to make sure this seems right. The equation indicates that the optimal level of labor demand is increasing in output prices, but decreasing in capital prices and wages.

Solve for the optimal input of capital by substituting the optimal level of labor demand.

\[ k^* = \left( \frac{1}{16} \frac{p^2}{r^{1/2}w^{3/2}} \right) \frac{w}{r} \]

Simplify,

\[ k^* = \frac{1}{16} \frac{p^2}{r^{3/2}w^{1/2}} \]

Again, check to make sure this seems right.

The optimal level of production is found by substituting optimal factor inputs into the production function:

\[ q^* = (h^*)^{1/4}(k^*)^{1/4} \]

Anything else can be found the same way (profits, total costs, marginal cost, etc.)
What is labor demand if $p = 100$, $r = 16$, and $w = 2$?

Simply substitute these prices into the labor demand equation:

$$h^* = \frac{100^2}{16^{1/2}2^{3/2}} = 55.243$$

(Note: there is some rounding error, so $h^*$ is approximately 55.243.)

*Are the First Order Conditions Satisfied?*

One easy way to check your work is to see whether the first order conditions are satisfied at these optimal levels of labor and capital demand.

*Calculating Labor Demand Elasticity with respect to Wages*

The labor demand elasticity with respect to wages is

$$\epsilon = \frac{\partial h^*(w, r, p)}{\partial w} \frac{w}{h^*}$$

For this example, the first part is

$$\frac{\partial h^*(w, r, p)}{\partial w} = \frac{1}{16} p^2 r^{-1/2} \left(-\frac{3}{2}\right) w^{-5/2} = \left(-\frac{3}{2}\right) \frac{p^2}{r^{1/2}w^{5/2}}.$$ 

Substituting this into the equation for $\epsilon$:

$$\epsilon = \left(-\frac{3}{2}\right) \frac{p^2}{r^{1/2}w^{5/2}} \frac{w}{h^*} = \left(-\frac{3}{2}\right) \frac{p^2}{r^{1/2}w^{3/2}} \frac{1}{h^*}$$
Notice that $\epsilon$ is a function of $w$, $p$, $r$, and $h^*$. (Note: this is how you should express $\epsilon$ when you are asked to express it as a function, i.e. do not substitute for the value $h^*$.)

By substituting for $w$, $p$, $r$, and $h^*$, we can calculate the exact elasticity of substitution at the point of optimal labor demand ($h^*$). (Note: in general, we could evaluate $\epsilon$ at other $h$ values, but these are not as relevant as the optimal labor demand point.)

Substituting for $w = 2$, $p = 100$, $r = 16$, and $h^* = 55.243$,

$$
\epsilon = \left(-\frac{3}{2}\right) \frac{100^2}{16^{1/2}2^{3/2}} \frac{1}{55.243} = -1.5
$$

(Note: there is some rounding error, so $\epsilon$ is approximately $-1.5$.)

Another way to find this number is to substitute the labor demand equation directly into the elasticity equation:

$$
\epsilon = \left(-\frac{3}{2}\right) \frac{p^2}{r^{1/2}w^{3/2}} \frac{1}{1/16} r^{1/2} w^{3/2}
$$

Simplifying,

$$
\epsilon = \left(-\frac{3}{2}\right) p^2 r^{-1/2} w^{3/2} 16 p^{-2} r^{1/2} w^{3/2} = \frac{-16 \times 3}{32} = -1.5.
$$

This value is the same as above.

This value of $\epsilon$ indicates that the labor demand elasticity with respect to
wages (at the point of optimal labor demand $h^*$) is elastic: $\epsilon > 1$.

### 3.10 Aggregate Labor Market

The total market demand for labor in the economy is the sum of the labor demand of each firm. If each firm $j$ has a different production function (indexed $j$), total labor demand is $\sum_{j=1}^{N} h_j^*(w, r, p)$, where $N$ is the number of firms. If all firms are identical, the market demand for labor is simply: $N * h^*(w, r, p)$, where $h^*(w, r, p)$ is the optimal labor demand of each firm.

(Note: I will use the same lower case $h$ to indicate market labor demand and individual firm labor demand. The distinction depends on the context.)

#### 3.10.1 What Affects Aggregate Demand for Labor?

As we have seen, several factors affect $h^*$ and the market demand for labor. The four most basic conclusions:

1) Lower wage rates increase $h^*$.

2) Higher output prices increase $h^*$.

3) Lower rental rates on capital can increase $h^*$ through a scale effect. A decline in $r$ reduces total costs and causes the firm to increase production. The increased level of production causes the firm to increase its labor demand.

4) Lower rental rates on capital can decrease $h^*$ through a substitution effect. As $r$ declines, capital becomes less expensive than labor.
3.10.2 Labor Demand Curve

The labor demand curve we discussed in the labor market overview section graphs the relationship between the aggregate $h^*$ for all firms and the wage rate ($w$). The other factors that affect $h^*$, such as $p$ and $r$, and the shape of the production function, affect the location of the curve (e.g. rotate it, shift it in or out). A given labor demand curve therefore reflects a particular production technology and prices ($p$ and $r$). Changes in $w$ are reflected in changes along a given labor demand curve.

3.10.3 Elasticity of Labor Demand and Market Demand

The elasticity of labor demand (with respect to wage rates) for the aggregate labor market depends on the $\epsilon$ value of individual firms. Consider the two extreme cases. Reality is somewhere in between these two cases.

*Perfectly Elastic Labor Demand*

If all firms have an $\epsilon = -\infty$, then the labor market demand curve is a flat (horizontal) line at the equilibrium wage rate ($w^*$). Labor demand is perfectly elastic in this case. An increase in the wage rate from the equilibrium rate causes all firms to lower their labor demand to 0.

*Perfectly Inelastic Labor Demand*

If all firms have an $\epsilon = 0$, then the labor market demand curve is a vertical line at the equilibrium number of hours used in the labor market ($h^*$). Labor
demand is perfectly inelastic in this case. An increase in the wage rate from the equilibrium rate does not change labor demand at all and simply raises the equilibrium wage rate.

### 3.11 Minimum Wage Laws

Minimum wages are price floors in the labor market. A minimum wage law states that no worker can be paid less than the minimum wage. Generally, minimum wage laws in the United States specify some exceptions to the law. Some workers in “uncovered” sectors are not subject to the minimum wage laws. Most workers are in the “covered” sector and must be paid at least the minimum wage. Minimum wages laws are set by all levels of government. Currently, the federal minimum wage is $5.15. Some states and city governments (mainly large cities and states where the cost of living is higher) set their own, higher, minimum wages. Some cities have passed “living wage” laws, which set even higher minimum wages.

See this website for some information on minimum wages in the United States:

http://www.dol.gov/esa/minwage/america.htm

### 3.11.1 Graphing a Minimum Wage

Figure 3 graphs a minimum wage law in a basic labor supply and demand graph. The minimum wage is set at $w$, which is above the equilibrium wage
level of $w^\ast$. A minimum wage set below the equilibrium wage would have no
effect. The minimum wage reduces equilibrium labor hours employed from
$h^\ast$ to $h'$. This reduction in employment because of the minimum wage is
often called the employment effect.

3.11.2 Labor Demand Elasticities and Minimum Wage

From our analysis of labor demand, we know that the demand for labor is
decreasing in wages (downward sloping labor demand curve). The extent of
employment effects depends on the labor demand elasticity with respect to
wage changes. Especially important to recognize is that the relevant labor
demand elasticity is that for the low wage labor market (e.g. manual laborers, workers in retail and fast food restaurants). How firms, which hire mainly workers with wages well above the minimum wage, respond to minimum wages is not relevant.

If the labor demand elasticity with respect to wages is low ($|\epsilon|$ is low),
the employment effects of minimum wages are low.

In the one extreme case, where labor demand is perfectly inelastic (a vertical labor demand curve), there are no employment effects. A minimum wage law in this case generates only the benefit of higher wages.

In the other extreme case, where labor demand is perfectly elastic (a horizontal labor demand curve), there are no wage gains resulting from a minimum wage wage. Faced with higher wage rates, firms with a perfectly elastic labor demand choose to hire no workers (e.g. they substitute capital
The reality is somewhere between the two extreme cases. The tradeoffs involved in minimum wage laws involve a choice between a labor market with higher employment and lower wages and a labor market with lower employment and higher wages.

Empirical research on this topic has generally found that the employment effects associated with recent minimum wage are small. This does not imply that minimum wage increases, especially large increases, would not have substantial employment effects. It is difficult to predict the effects of minimum wage changes because of the difficulty of estimating labor supply and demand functions. Most empirical work which analyzes minimum wages examines the effects of prior changes. These previous experiences may have only limited relevance to future, higher changes.
4 Labor Supply

4.1 Model of Labor Supply

The following is a simple model of individual labor supply decisions. At the end of this section, we will examine extensions to this model to make it more realistic.

In this model, individuals choose how many hours to work in the labor market \((h)\) and how many consumer goods to purchase \((c)\). Every hour the individual does not work is leisure time \((l)\). Leisure is essentially another good that the individual can purchase by not working. The total amount of time an individual has is \(T\). Therefore, the *time constraint* in the problem is \(T = l + h\). For every hour the individual works, she receives a wage \(w\) per hour.

Individuals have preferences over the two goods, consumer goods \((c)\) and leisure \((l)\). Preferences are indicated by a utility function: \(U(c, l)\). The utility function provides the number of “utils” or satisfaction the individual receives from different combinations of \(c\) and \(l\).

The price for consumer goods is \(p\). The price of leisure is foregone wages or \(w\). Like our labor demand model, this is a *partial equilibrium* model and these prices are taken as given by the individual. The prices are *exogenous*.

The individual is assumed to have two sources of income: labor income from working \((wh)\) and non-labor income called \(V\). \(V\) can be thought of as wealth the individual has accumulated (e.g. savings or inheritance from
a rich relative). The individual receives this non-labor income even if she chooses not to work \((h = 0)\).

### 4.2 Utility Maximization

The individual maximizes her utility by choosing \(c\) and \(l\) subject to a budget constraint and a time constraint. The budget constraint is that total income must equal total expenditures on consumer goods:

\[
V + wh = pc. 
\]

Total income is non-labor income plus labor income \((V + wh)\). Total expenditure on consumer goods is \(pc\).

The maximization problem is then

\[
\max_{c,l} U(c, l) \quad s.t. \quad V + wh = pc \text{ and } T = l + h. 
\]

Re-written in terms of leisure,

\[
\max_{c,l} U(c, l) \quad s.t. \quad V + w(T - l) = pc 
\]

Without a more complex setup (i.e. writing the constrained optimization problem as a Lagrangian function) or specifying the functional form of the utility function (see the Cobb-Douglas example below), we cannot derive the
first order conditions explicitly, as in the labor demand analysis. Instead, we will solve this optimization problem by stating that the optimal leisure hours choice and consumer goods choice must satisfy this *tangency condition*:

$$\frac{\partial U(c, l)}{\partial l} = \frac{w}{p} \frac{\partial U(c, l)}{\partial c}.$$

Below, we will describe how this tangency condition relates to a graphical representation of the utility maximization problem.

With $w$, $p$, $V$, $T$, and a functional form for the utility function, we can solve this problem for the optimal number of leisure hours demand $l^*(w, p, V)$, the optimal purchase of consumer goods $c^*(w, p, V)$, and the optimal labor supply decision $h^*(w, p, V) = T - l^*(w, p, V)$.

### 4.3 Some Terminology

**Marginal Utility of Leisure**

$$MU_l = \frac{\partial U(c, l)}{\partial l}$$

**Marginal Utility of Consumption**

$$MU_c = \frac{\partial U(c, l)}{\partial c}$$

**Marginal Rate of Substitution**
\[ MRS = \frac{MU_l}{MU_c} \]

This is the \( MRS \) of \( l \) for \( c \).

4.4 What Does the Tangency Condition Mean?

The marginal utility of leisure provides the value of one more hour of leisure. The cost of obtaining this extra leisure is that the individual must give up an hour of work, which would provide \( w \) of additional income. \( w \) therefore is the “price” of leisure. \( w \) is the marginal cost of leisure.

The marginal utility of consumer goods provides the value of one more unit of consumer goods. The marginal cost of obtaining an additional unit of consumer goods is the price of the consumer goods \( p \).

The intuition behind the first order conditions is that the individual should continue to purchase leisure and consumption goods until the point at which the ratio of the marginal benefit of the two goods (\( MRS \)) equals the ratio of the marginal costs (\( w/p \)). The tangency condition is simply

\[ MRS = \frac{w}{p} \]

This tangency condition is simply a re-statement of a fundamental principle of economics: decision makers make choices by equating the marginal
benefit with the marginal cost.

4.5 Labor Supply Elasticity with Respect to Wages

The labor supply elasticity with respect to wages is defined as:

$$\gamma = \frac{\%\Delta h^*(w, p, V)}{\%\Delta w},$$

or

$$\gamma = \frac{\partial h^*(w, p, V)}{\partial w} \cdot \frac{w}{h^*}.$$

Or in logs,

$$\gamma = \frac{\partial \ln h^*(w, p, V)}{\partial \ln w}.$$

The labor supply elasticity with respect to wages can be positive or negative: $-\infty \leq \gamma \leq +\infty$.

The labor supply elasticity has the following interpretation. This interpretation and terminology is similar to that for the labor demand elasticity. However, note that $\gamma$ can be negative or positive, so the interpretation is for the absolute value of $\gamma$: $|\gamma|$.

i) $|\gamma| = 0$ (*perfectly inelastic* labor supply). Changes in wage rates do not affect labor supply.
ii) $0 < |\gamma| < 1$ (inelastic labor supply). If wages change by X %, labor supply changes by less than X %.

iii) $|\gamma| = 1$ (unit elastic labor supply). If wages change by X %, labor supply changes by exactly X %.

iv) $|\gamma| > 1$ (elastic labor supply). If wages change by X %, labor supply changes by more than X %.

v) a) $\gamma = \infty$ (perfectly elastic labor supply). If wages change by X %, labor supply increases to $\infty$.

v) b) $\gamma = -\infty$ (perfectly elastic labor supply). If wages change by X %, labor drops to 0.

It is important to note that $\gamma$ is (in general) a function of prices ($w$ and $p$), level of labor hours or leisure ($h$ or $l$), and consumer goods ($c$). This means that the labor supply elasticity can vary as these factors change.

### 4.6 Income and Substitution Effects

We can decompose how changes in wage rates affect labor supply into two factors. Consider an increase in wages.

*Substitution Effect*

Higher wages increase the return to work and raise the price of the leisure good. The worker substitutes away from leisure toward consumer goods. This causes the individual to demand less leisure and work more.
**Income Effect**

Higher wages increase the labor income for the individual. In effect, the individual is wealthier. If leisure is a *normal* good (demand for leisure is increasing in income), then higher income causes the individual to consume more leisure and work less.

**Key**

Unlike the scale and substitution effects with labor demand, the substitution and income effects in the labor supply case move in the opposite direction. Which effect is stronger determines whether an increase in wage rates reduces or increases labor supply.

If the substitution effect is larger than the income effect, an increase in wages increases labor supply.

If the income effect is larger than the substitution effect, an increase in wages decreases labor supply.

### 4.7 Indifference Curves and Budget Lines Graphs

Figure 4 graphs the utility maximization problem. On the vertical axis is units of consumption \(c\). On the horizontal axis are hours of leisure \(l\). The two objects in the graph are an indifference curve and a budget line.

**Budget Line**

The budget line is the budget constraint from above
\[ V + wh = pc. \]

Re-written in terms of leisure, the budget line becomes

\[ V + w(T - l) = pc. \]

Re-arranging for an equation of a line \((y = c \text{ and } x = l)\).

\[ c = \frac{V}{p} + \frac{w}{p} T - \frac{w}{p} l \]

The vertical intercept is \(c = \frac{V}{p} + \frac{w}{p} T\). If the individual always works (consumes 0 leisure), she receives \(wT\) in labor income. Therefore, she can buy \(\frac{V}{p} + \frac{w}{p} T\) worth of consumer goods.

In this simple model, \(\frac{w}{p}\) is the real wage. As opposed to \(w\) which is the nominal wage.

Notice that the horizontal intercept \((c = 0)\) is never reached. Even if the individual never works \((l = T)\), she still can use her non-labor income \(V\) to purchase \(\frac{V}{p}\) units of consumer goods.

The slope of the line is the negative of the price ratio \(-\frac{w}{p}\). This indicates the rate at which workers can trade leisure for consumer goods.

*Indifference Curve*
The indifference curve is determined by the utility function. It indicates the preferences the individual has over consumer goods and leisure. Along an indifference curve, utility is constant.

The equation of an indifference curve is

$$U_0 = U(c, l),$$

where $U_0$ is some constant utility. The slope of the indifference curve is the marginal rate of substitution between leisure and consumer goods.

Indifference curves can take any of the forms of production functions discussed above: perfect substitutes, perfect complements, or Cobb-Douglas. The analysis of each of these types of production functions follows through to the utility functions.

For example, a Cobb-Douglas utility function would take the form

$$U(c, l) = c^{\theta_1} l^{\theta_2}$$

*Tangency Condition*

Where the indifference curve and the budget line are tangent gives the optimal combination of consumption and leisure. At this point, the $MRS = w/p$. This is simply a graphical representation of tangency condition we discussed above.
4.8 A Cobb-Douglas Example

An example: \( U(c, l) = cl \).

Marginal utility of consumption:

\[
MU_c = l
\]

Marginal utility of leisure:

\[
MU_l = c
\]

Utility maximization problem is

\[
\max_{c,l} cl \quad s.t. \quad V + w(T - l) = pc.
\]

We noted above that this condition always holds:

\[
MRS = \frac{w}{p}.
\]

In this problem,

\[
MRS = \frac{MU_l}{MU_c} = \frac{c}{l}.
\]

Substituting,
\[
\frac{c}{l} = \frac{w}{p}
\]

Solve for \( c \),

\[
c = \frac{lw}{p}
\]

Now substitute this into the budget constraint:

\[
\frac{lw}{p} = \frac{V}{p} + \frac{w}{p}(T - l)
\]

Simplify,

\[
lw = V + w(T - l)
\]

\[
l = \frac{V}{w} + T - l
\]

\[
2l = \frac{V}{w} + T
\]

\[
l^*(w, p, V, T) = \frac{1}{2}\left(\frac{V}{w} + T\right)
\]
Check to see if this seems right. Leisure consumption is decreasing in its price, $w$. Leisure consumption is increasing in non-labor income $V$.

*Another Way to Solve this Problem*

Another way to solve this problem is to substitute for consumption from the budget constraint: $c = \frac{V}{p} + \frac{w}{p}(T - l)$. This reduces the maximization problem above to a choice of leisure only.

$$\max_l \left[ \left( \frac{V}{p} + \frac{w}{p}(T - l) \right) l \right]$$

Simplify,

$$\max_l \left[ \frac{V}{p} l + \frac{w}{p} T l - \frac{w}{p} l^2 \right]$$

The first order condition with respect to leisure is

$$\frac{V}{p} + \frac{w}{p} T = 2 \frac{w}{p} l$$

Solve for optimal $l$,

$$l^*(w, p, V, T) = \frac{1}{2} \left( \frac{V}{w} + T \right)$$

This is the same equation as above.

*Continuing the Problem*
With an equation for optimal leisure demand, we can find everything else.

Optimal labor supply is simply

\[ h^*(w, p, V, T) = T - l^*(w, p, V, T) = T - \frac{1}{2} \left( \frac{V}{w} + T \right). \]

Simplifying,

\[ h^*(w, p, V, T) = \frac{1}{2} T - \frac{1}{2} \frac{V}{w} \]

If the person has no non-labor income \((V = 0)\), she works exactly 1/2 of her time, regardless of the wage. This is because of the particular utility function we assumed in which leisure and consumer goods are equally valued.

We can also calculate the optimal consumption of consumer goods using the optimal leisure decision.

\[ c^*(w, p, V) = \frac{V}{p} + \frac{w}{p} [T - \frac{1}{2} (\frac{V}{w} + T)] \]

Simplifying,

\[ c^*(w, p, V) = \frac{1}{2} \frac{V}{p} + \frac{1}{2} \frac{w}{p} T \]

Check to see if this seems right. Consumption is increasing in non-labor income. Consumption is decreasing in the price of consumer goods. Consumption is also increasing in the wage rate because of an income effect.

Note that if there is no non-labor income, the individual can buy \( \frac{1}{2} w T \)
worth of consumer goods.

What is $l^*$, $h^*$, and $c^*$ with $p = 2$, $V = 8$, $w = 4$?

To find the exact values of $l^*$, $h^*$, and $c^*$ given these prices, simply substitute the prices into each of our equations. Assume the individual has $T = 16$ total hours each day to devote to leisure or working (i.e. the individual sleeps for 8 hours each day).

$$l^* = \frac{1}{2}(\frac{8}{4} + 16) = 9.$$  

Check to make sure this is feasible: $l^* < T$ and $9 < 16$.

(Note: It is possible that as non-labor income increases to very high numbers (e.g. $V = 1000$), the optimal leisure decision will exceed the amount of time available: $l^* > T$. This indicates that the person is so wealthy that she chooses to never work.)

Optimal labor supply is

$$h^* = T - l^* = 16 - 9 = 7.$$  

Or using the equation we derived above.

$$h^* = \frac{1}{2}T - \frac{1}{2} \frac{V}{w} = \frac{1}{2} \cdot 16 - \frac{18}{24} = 8 - 1 = 7.$$  

Finally, optimal consumption of consumer goods is
\[ c^* = \frac{18}{2} + \frac{14}{2} \cdot 16 = 2 + 16 = 18. \]

**Is the Time Constraint Satisfied?**

This time constraint must hold at the values of labor supply and leisure demand we calculated.

\[ T = h^* + l^* \]

\[ 16 = 7 + 9 \]

\[ 16 = 16 \]

**Is the Budget Constraint Satisfied?**

Check to make sure the budget constraint is satisfied:

\[ V + w(T - l) = pc \]

\[ 8 + 4(16 - 9) = 2 \cdot 18 \]
Is the Tangency Condition Satisfied?

The tangency condition must hold at the values of \( l^* \) and \( c^* \) we calculated.

\[
MRS = \frac{w}{p}.
\]

In this problem,

\[
MRS = \frac{MU_{l^*}}{MU_{c^*}} = \frac{c^*}{l^*} = \frac{18}{9}.
\]

The price ratio is

\[
\frac{w}{p} = \frac{4}{2}.
\]

Therefore this condition holds:

\[
\frac{18}{9} = \frac{4}{2}.
\]

Labor Supply Elasticity with Respect to Wages

The labor supply elasticity with respect to wages is defined as
\[ \gamma = \frac{\partial h^*(w, p, V)}{\partial w} \frac{w}{h^*}. \]

For our example, the first part is

\[ \frac{\partial h^*(w, p, V)}{\partial w} = -\frac{1}{2} \left(-1\right) \frac{V}{w^2} = \frac{1}{2} \frac{V}{w^2} \]

Substituting,

\[ \gamma = \frac{1}{2} \frac{V}{w^2} \frac{w}{h^*} = \frac{1}{2} \frac{V}{w h^*}. \]

This is the labor supply elasticity expressed as a function (at the point of optimal labor supply \( h^* \)).

Given our prices and non-labor income, \( p = 2, V = 8, \) and \( w = 4, \) and the optimal labor supply value calculated above \( h^* = 7, \) the elasticity of labor supply is

\[ \gamma = \frac{18}{247} = \frac{1}{7}. \]

At this point, labor supply is relatively inelastic.

Another way to calculate the labor supply elasticity is to substitute the labor supply function:
\[ \gamma = \frac{1}{2} \frac{V}{w} * \frac{1}{2} \frac{T}{w} - \frac{1}{2} \frac{V}{w}. \]

Simplifying,

\[ \gamma = \frac{1}{2} \frac{V}{w} 2* \frac{1}{T} \frac{V}{w} = \frac{V}{w} * \frac{1}{T - \frac{V}{w}} \]

Substituting values,

\[ \gamma = \frac{8}{4} * \frac{1}{16} - \frac{8}{4} = 2 * \frac{1}{14} = 1/7. \]

Same value as above.

### 4.9 Labor Supply Curve

Like the labor demand curve, the labor supply curve is the sum of the individual labor supplies.

If each individual \( i \) has different preferences \( \sum_{i=1}^{N} h_i^*(w, p, V) \). If all individuals are identical, then labor supply is simply: \( N * h^*(w, p, V) \)

As we have seen, several factors affect \( h^* \) and the market supply of labor.

The three most basic conclusions:

1) Higher non-labor income leads to lower labor supply through an income effect.
2) If the substitution effect is larger than the income effect, higher wages increase labor supply.

3) If the income effect is larger than the substitution effect, higher wages decrease labor supply.

It may be the case that these two wage effects have varying strengths over the range of possible wages. Figure 5 graphs a *backward bending* labor supply curve. Over the range of wages from \( w = 0 \) to \( w = w' \), higher wages are causing the workers to supply more labor. In this area, the substitution effect is dominating the income effect. At wages \( w' \) and higher, the income effect is larger than the substitution effect. At this point, higher wages are causing the worker to work less and consume more leisure.

### 4.10 Reservation Wages and the Decision to Work

The models examined thus far assume workers smoothly adjust their hours of work in response to changes in wage rates or non-labor income. Here we consider a model where workers can decide whether to work at all. The basic model assumes that there is some reservation wage that the worker would need to be offered before she works. If she is offered a wage below her reservation wage, she refuses to work. If she is offered a wage above her reservation wage, she works at that wage.

Call the reservation wage, \( w \). If \( w > w \), then the worker works. If \( w < w \), the worker works zero hours. If \( w = w \), the worker is just indifferent between
working and not.

Figure 6 displays a reservation wage model. There are two budget lines corresponding to different wages (remember that the slope of budget line is $-w/p$). On the budget line corresponding to the reservation wage, the indifference curve is tangent at the point of $l = T$. Here the worker is just indifferent between working and not. At the lower budget line, with wage $w < w^*$, there is no leisure consumption level, $l < T$, along this budget line where the worker receives as much utility as she receives at $l = T$. Therefore, it is not optimal for the worker to work at all if she is offered a wage less than the reservation wage. At wages higher than reservation wage (these budget lines are not drawn in Figure 6), it is optimal for the worker to work and choose $l < T$.

4.11 Welfare Programs

4.11.1 Cash Grants

There are many variants of welfare programs. One basic welfare program provides a cash grant to individuals who do not work. (To simplify the model, we will assume the individual has the option of working. This model would not apply to individuals who are disabled and cannot work.)

A cash grant for non-working individuals changes the budget line the individual faces. Thus far, our budget lines have this form:
\[ c = \frac{V}{p} + \frac{w}{p} T - \frac{w}{p} l \]

For simplicity, assume that the individual has no non-labor income \((V = 0)\).

Re-write the budget line:

\[ c = \frac{w}{p} T - \frac{w}{p} l. \]

The welfare program rules state that if an individual does not work \((l = T)\), then she receives \(M\) dollars. If she works even one hour, she receives no welfare assistance \((a \text{ take it or leave it program})\). With these welfare rules, the budget line has two parts:

\[ c = \frac{w}{p} T - \frac{w}{p} l \quad \text{if} \quad l < T, \]

and

\[ c = \frac{M}{p} \quad \text{if} \quad l = T. \]

Figure 7 graphs this form of budget line. The budget line is in bold and has slope \(-w/p\) until \(l = T\). At \(l = T\), the budget line is a single point \(c = M/p\). In this graph, without the cash grant welfare program, the
individual choose to work $T - l^*$ hours and consume $c^*$ consumer goods. With the welfare program, the individual receives higher utility by not working, accepting the cash grant, and consuming $c = M/P$ consumer goods.

In this model, the welfare program has reduced labor supply. The extent to which a welfare program reduces labor supply depends on the size of the cash payment and an individual’s preferences for leisure and consumption (the shape of the indifference curve). As long as leisure is a normal good, it is not possible for a welfare program to increase labor supply.

4.11.2 Earned Income Tax Credit

An alternative type of welfare program attempts to provide incentives for working by increasing the effective wage for low income people. The Earned Income Tax Credit (EITC) is one such program in the United States. This program provides individuals a tax credit based on the income they earn through working. One reason to prefer this type of welfare program over a cash grant program is that it can provide both income assistance for low wage workers and increase labor supply.

Here is a model of an EITC-like program. Workers pay no taxes, but receive a wage subsidy per each hour they work. If the worker works at least $h = T - l_a$ hours, they receive a wage of $(1 + \alpha)$ times the market wage $w$. For example, if the market wage is $10 and $\alpha$ is 0.1, the worker would receive $11 per hour under the EITC. $\alpha$ can be negative as well if taxes are decreasing the wage rate (e.g. $\alpha = -0.1$ means the worker receives 0.9 dollars
for every 1 dollar in wages). If the worker works between $h = T - l_a$ and $h = T - l_b$ hours, then she receives the market wage. If the worker works less than $h = T - l_b$ hours, she receives a wage of $(1 + \beta)$ times the market wage.

Like a cash grant welfare program, our EITC-like program changes the budget line. The usual budget line (with $V = 0$) is

$$c = \frac{w}{p} (T - l).$$

Our hypothetical EITC rules transform the budget line into three sections.

$$c = (1 + \alpha) \frac{w}{p} (T - l) \quad \text{if} \quad 0 < l < l_a,$$

and

$$c = \frac{w}{p} (T - l) \quad \text{if} \quad 1_a < l < l_b,$$

and

$$c = (1 + \beta) \frac{w}{p} (T - l) \quad \text{if} \quad 1_b < l < T.$$

This EITC-like program is graphed in Figure 8. With no EITC policy, the individual consumes $l^*$ leisure.

Our EITC rules transform the budget line into a *kinked* budget line.
Under the EITC rules, the individual increases their consumption of leisure to $l'$. This reduction in labor supply as result of the EITC like program indicates that the income effect of higher wages is larger than the substitution effect.

This is not the only outcome from an EITC program. The result that EITC reduces labor supply is a direct result of the assumptions about the utility function and the shape of indifference curves. Under different assumptions, an EITC-like program can increase labor supply or have no effect.

Whether the EITC reduces labor supply is an empirical question. Most research has found that the EITC in the United States increases labor supply and causes more people to enter the labor market. We will examine the methodology underlying these findings later in the course.
5 Equilibrium and Market Structure

5.1 Equilibrium in the Labor Market

In the overview of the labor market section, we briefly described an equilibrium in the labor market. An equilibrium is reached when the labor demanded by firms is equal to the labor supplied by workers. This is the point of intersection of the labor supply and labor demand curves. At this point, we have an equilibrium number of labor hours employed in the economy $h^*$ and an equilibrium wage rate $w^*$. The equilibrium condition can be summarized as

$$h^*_s(w^*) = h^*_d(w^*),$$

where $h^*_s(w^*)$ is the optimal labor supply (of all workers) at the wage $w^*$, and $h^*_d(w^*)$ is the optimal labor demand (of all firms) at the wage $w^*$.

5.2 Out of Equilibrium

Let’s consider what happens when the labor market is not at an equilibrium. Consider two cases.

Case 1 Labor Surplus

The wage in the labor market is higher than the equilibrium wage $w > w^*$. At a wage of $w > w^*$, the optimal labor supply is greater than the optimal
labor demand. There is a surplus of labor in the economy at this $w$:

$$h^*_s(w) > h^*_d(w).$$

At this wage $w$, there are more people willing to work than there are jobs. Some of these people may be unemployed and others may be working fewer hours than they would like to.

**Case 2 Labor Shortage**

The wage in the labor market is lower than the equilibrium wage $w < w^*$. At a wage of $w < w^*$, the optimal labor supply is less than the optimal labor demand. There is a shortage of labor in the economy at this $w$:

$$h^*_s(w) < h^*_d(w),$$

At this wage $w$, there are too few people willing to work. There are some jobs which are not filled. Some jobs are filled, but the workers are not willing to work more hours (e.g. overtime) at this lower than equilibrium wage $w < w^*$.

### 5.3 Reaching a Labor Market Equilibrium

How does the economy move from out of equilibrium (either a labor surplus or shortage) to a labor market equilibrium?
Consider the two cases again.

Case 1 *Labor Surplus*

In this case the prevailing wage $w$ is greater than the equilibrium wage $w^*$. With this excess supply of labor, workers are competing with each other for a limited number of jobs. There is an excess number of workers lined up outside a firm’s door. The process of reaching an equilibrium can be thought of in two ways: 1) as the workers compete with each other, they offer to work at a lower wage rates, or 2) the firm offers a lower wage rate until the excess supply of labor disappears. That is, as the wage rate falls, some workers no longer are willing to work at that lower wage. The wage rate $w$ falls until $w = w^*$ and the labor surplus disappears:

$$h_s^*(w^*) = h_d^*(w^*).$$

Case 2 *Labor Shortage*

In this case the prevailing wage $w$ is less than equilibrium wage $w^*$. With this shortage of labor, firms are competing with each other for a limited number of workers. There are a number of firms (employers) lined up outside each worker’s door. The process of reaching an equilibrium can be thought of in two ways: 1) as the firms compete with each other, they offer workers higher wage rates, or 2) the workers ask for a higher wage rate until the shortage of labor disappears. That is, as the wage rate rises, some firms no
longer want to employ labor at that higher wage. The wage rate $w$ rises until $w = w^*$ and the labor shortage disappears:

$$h^*_s(w^*) = h^*_d(w^*).$$

### 5.4 Arbitrage Across Labor Markets

What we described above is the process by which a single labor market reaches an equilibrium point. Now, let’s consider what happens to the equilibrium point in two separate labor markets.

#### 5.4.1 Two Geographically Separated Labor Markets

Consider two geographically separated labor markets. One is the United State labor market and the other is the Chinese labor market.

**Chinese Labor Market**

The supply of labor in China (from Chinese workers) and demand for labor in China (from Chinese firms) yields an equilibrium wage rate in China of $w^*_{ch}$ and equilibrium number of labor hours in China of $h^*_{ch}$.

**United States Labor Market**

Similarly, the supply of labor in United States (from US workers) and demand for labor in the US (from US firms) yields an equilibrium wage rate in the US of $w^*_{us}$ and equilibrium number of labor hours in the US of $h^*_{us}$. 
Assume (as is the case in reality) that the US equilibrium wage rate is higher than the Chinese wage rate:

\[ w_{us}^* > w_{ch}^*. \]

What do workers and firms do in each country?

**Firms**

Because wage rates are lower in China, US firms have an incentive to move to China and produce their goods in China. As US firms move to China, the labor demand for Chinese workers increases (a shift out of the Chinese labor demand curve). This produces an increase in the equilibrium Chinese wage rate.

**Workers**

Because wage rates are higher in the US, Chinese workers have an incentive to emigrate to the United States and work in the US. As Chinese workers emigrate to the US, the labor supply of US resident workers (immigrants and natives) increases (a shift out of the US labor supply curve). This produces a decrease in the equilibrium US wage rate.

**Arbitrage**

This movement of workers to higher wage labor markets (from China to the US) and firms to lower wage labor markets (from US to China) causes
the equilibrium wage rates in the two countries to converge. This general process is often called *arbitrage*. If the arbitrage is complete, the wage rate in China and the wage rate in the US will be equal:

\[ w_{us}^* = w_{ch}^* \]

Although these two labor markets are geographically separated, the mobility of firms and workers implies that the labor demand and supply in each labor market interact and affect each country’s respective equilibrium wage rate.

### 5.4.2 Arbitrage Across Skill Levels

Consider another example. There are two types of labor defined by the level of skill. One type of labor is college educated labor provided by workers with a college degree. The other type of labor is non-college educated labor provided by workers who do not have a college degree. The two types of labor work in the same geographic area (e.g. same city), but for each type of labor there is a separate labor market.

*Labor Market for College Education Workers*

The supply of college educated labor (from college graduates) and the demand for college educated labor (from firms that employ college educated labor) yields an equilibrium wage rate for college educated labor of \( w_c^* \) and
equilibrium number of college educated labor hours \( h^*_c \).

**Labor Market for Non-College Educated Workers**

The supply of non-college educated labor (from workers who do not have a college degree) and the demand for non-college educated labor (from firms that employ non-college educated labor) yields an equilibrium wage rate for non-college educated labor of \( w^*_{nc} \) and equilibrium number of non-college educated labor hours \( h^*_{nc} \).

Assume (as is the case in reality) that the equilibrium wage rate for college educated labor is higher than the wage rate for non-college educated labor:

\[
w^*_c > w^*_{nc}.
\]

**What do workers and firms in each labor market do?**

**Firms**

Because college educated labor is more expensive relative to non-college educated labor, firms have an incentive to substitute non-college educated labor for college education labor. This reduces the demand for college educated labor (a shift in of the demand curve in this market) and increases the demand for non-college educated labor (a shift out of the demand curve in this market). This shift in demand, decreases the college educated wage rate and increases the non-college educated wage rate.
Workers

Because the wage rate is higher for college educated workers, there is an incentive for non-college educated individuals to attend college and become college educated workers (i.e. enter the college educated labor market). This shifts the supply curve of college educated workers out and shifts in the supply curve of non-college educated workers. This shift in of supply decreases the college educated wage rate and increases the non-college educated wage rate.

Arbitrage

This movement of workers to higher wage occupations (from non-educated to educated) and firms to lower wage labor (from educated to non-educated) causes the equilibrium wage rates in the two labor markets to converge. If the arbitrage is complete, the wage rate for both types of labor will be equal:

\[ w_c^* = w_{nc}^*. \]

5.5 Labor Market Frictions

In reality, the wage rate has not fully converged in either example. The wage rate in China and the United States is not equal and college graduates are paid more than non-college graduates.

Labor market frictions are the explanations for the failure of wage rates to converge across all labor markets. A labor market friction is some cost
to arbitrage across labor markets. If there are no labor market frictions, we would expect the wage rate to be the same across all labor markets.

Types of Labor Market Frictions:

1) Mobility Costs

These are the costs of physically moving production to a new location (the costs of moving a US firm to China) or the costs of workers moving to a new labor market (the costs for Chinese workers emigrating to the US). For some firms, such as those in manufacturing, the cost of moving to a new labor market is relatively small. For other firms, such as firms in service industries, these costs are relatively larger. It would be prohibitively costly for my local grocery store to move to China, use cheaper Chinese labor, and ship my groceries back to me in the US. However, we have recently seen that some services, such as telemarketing or phone help, are being “out-sourced” to countries like India where labor costs are lower.

2) Search Costs

These are the costs borne by firms in finding new workers or the costs borne by workers in finding new employers. There are some costs incurred during the process of searching for a good match between employers and employees. For workers, search costs can include the costs of walking around and interviewing at different jobs. For firms, the costs can include the costs
of using employment agencies, headhunters, and interviewing and screening job applicants.

3) **Cost of Human Capital Investment**

Both firms and workers may need to pay for the cost of new human capital as firms and workers enter new labor markets. For example, US firms moving to China may have to invest in training their new Chinese workers. Chinese workers moving to the US will incur some costs in learning English or other new skills. Workers who want to move from the non-college educated labor market to the college educated labor market incur the costs of time and tuition in obtaining a college degree.

Some forms of human capital are impossible to obtain because the human capital stems from skills or talents an individual is born with. The costs for workers to acquire these skills if they are not born with them is prohibitively high. For example, professional basketball players earn much more than I do and I would like to move into their labor market. However, I was not born 7 feet tall or with any athletic ability. The cost for me to obtain these types of human capital and enter the professional basketball labor market is essentially infinity.

4) **Institutional Rules**

There may also be a number of institutional rules which create labor market frictions and increase the cost of arbitrage across labor markets. Many
governments have explicit rules that impose costs on firms looking to move to new labor markets or hire or fire workers. If there is a government rule that states a firm must pay a severance package to laid off workers or contribute to an unemployment insurance fund, this rule imposes costs on firms who would like to replace their current workers or move to another labor markets. In labor contracts negotiated by unions, restrictions may be placed on the extent to which a firm can lay off workers or hire new workers.

5.6 Market Structure

Frictions in the labor market change the structure of labor markets. A labor market without frictions, in which firms and workers can costlessly enter the labor market, are called *competitive* labor markets. These markets are often characterized by a large number of firms or workers in the market. A *non-competitive* market, on the other hand, may have fewer firms or workers because of the presence of some labor market friction that imposes costs to entry into the market.

*Types of Labor Market Structures*

*Competitive Labor Market*

These types of markets are typically characterized by a large number of firms (employers) and workers. The firms have a large pool of workers to choose from. And the workers have a large number of employers to choose
from. This creates competition for jobs among workers and competition among firms for workers. No firm or worker has any market power and therefore cannot influence the market price of labor, the wage rate.

**Monopsony in the Labor Market**

If there is only one firm in a labor market, this firm is called a monopsonist. The one firm is the one employer or consumer of labor. A monopsonist could be the one employer in a geographic area (e.g., the one company town). Or, the monopsonist firm could be the one employer of a particular occupation or set of skills (e.g., Major League Baseball is the only American employer in the market for professional baseball players). In addition, groups of independent firms may join together in the form of a cartel to act jointly as a monopsonist employer.

Typically, monopsonist firms employ less labor (lower labor demand) than a competitive firm. The lower labor demand pushes the equilibrium wage rate in a labor market with a monopsonist firm lower than the equilibrium wage rate in a competitive labor market. In addition, monopsonist firms typically earn higher profits than firms in a competitive labor market. To maintain the market power of the monopsonist, there must be some sort of labor market friction which prevents other firms from entering this labor market.

**Monopoly in the Labor Market**

If there is only one worker in a labor market, the one worker is a monopolist in this labor market. Although true monopoly labor markets are
probably non-existent, many workers enjoy some level of market power in their labor market. An example would be entertainers or athletes for which there are only limited or imperfect substitutes.

Typically, monopolist workers restrict the amount of their labor supply. Just like a monopolist firm in the product market, this supply restriction increases the equilibrium wage monopolist worker receives relative to a competitive labor market. If there were many good substitutes for certain entertainers or athletes, their market power would dissipate and the wages they receive would fall. There must be some sort of labor market friction that prevents individuals from entering these markets and capturing the rents the monopolist workers receive. For example, my lack of talent prevents my entry into the basketball labor market.

A Bilateral Monopoly Problem

If there is both a single monopsonist firm and a single monopolist worker, the market structure is called a bilateral monopoly. In this labor market, both the firm and the worker have market power. How the equilibrium wage rates and labor hours are determined in this market is an open question.

Monopoly in the Product Market

If there is a monopoly in the product market, there is only one firm that produces a particular product. This does not imply anything necessarily about the structure of labor market the monopolist firm is in. Classic examples of firms with substantial market power in the product market are the
OPEC (Organization of Petroleum Exporting Countries) cartel, which is a cartel of oil producers, and the firm DeBeers, which owns most of the world’s supply of diamonds.

Monopolist firms restrict output and raise output prices relative to firms in a competitive product market. These monopolist firms often earn higher profits than firms in competitive output markets. An interesting question for the labor market is who gets to keep these rents: labor or capital. The monopolist firm could return at least part of these rents to the workers in form of higher wages or give them to the owners of the capital (the firm’s owners). Because of the presence of these rents, the employees of monopolist firms may have a greater opportunity to extract higher wages from monopolist firms than from firms in competitive product markets.
6 Unions

6.1 Basic Facts about Unions

Unions are collections of workers. There are three main activities of unions: collective bargaining with employers over employment contracts, providing social services to union members (e.g. job training), and political lobbying. About 13 percent of workers in the United States belong to unions. This is a decline from a peak of about 25 percent in the 1950-70s period. In Western Europe, far more workers belong to unions, and the unions there are more involved in politics. The main focus of unions in the United States is collective bargaining. Most union members are blue collar workers in the construction, manufacturing, and transportation industries. More recently, unions have increased their membership among public sector/government employees: police and firemen, teachers, postal workers. About 40 percent of public sector employees are union members.

Unions are formed after a majority of workers at firm vote in an election to certify a particular union as their collective bargaining representative. This gives the union the sole right to bargain over labor contracts. The union negotiated contracts apply to all workers, regardless of their individual union membership. In some states, all workers in a unionized firm are required to join the union. In 22 states, “right-to-work” legislation allows workers to work at a unionized firm without joining the union. The non-union workers in unionized firms have varying rights to negotiate their own labor contracts.
separately with the employer.

6.2 What Do Firms and Unions Bargain Over?

Unions and firms can potentially bargain over all aspects of the employment contract, including

1) Wage rates: starting salaries, criteria for salary increases (e.g. cost of living adjustments), overtime pay.

2) Benefits: pension and health insurance.

3) Rules regarding hiring, both the numbers of new hires and the selection criteria for hiring.

4) Rules regarding promotion within the firm.

5) Rules regarding firing and layoffs; what are grounds for dismissal; and who gets laid off first.

6) Level and types of training workers receive from the firm.

7) Work and safety conditions.

6.3 Collective Bargaining

The process of collective bargaining over labor contracts can take several forms.
i) Agreement

The employer and the union representatives come to an agreement regarding the labor contract. Depending on the union rules, the labor contract may need to be approved by a vote of the union membership.

ii) Mediation

If the employer and the union representatives cannot come to an agreement, some sort of mediation may be used. This can take the form of formalized arbitration in which an independent, and hopefully objective, individual (an arbitrator) helps the employer and union representation reach a compromise labor contract. Prior to the start of negotiations, the employer and union representative may agree to some form of arbitration which will take place if the parties cannot come to an agreement on their own. If the employer and union agree to binding arbitration, the ruling of the arbitrator must be abided by the employer and union.

iii) Strikes and Lockouts

If an agreement on a labor contract cannot be reached, the workers may strike and withhold all labor services. Likewise, the employer can lockout the workers, shut down production, and withhold all employment. Often, workers are simultaneously on strike and are locked out by employers. The difference is trivial.
Both sides lose from either a strike or lockout. Workers lose wages and firms lose profits. Strikes and lockouts can also be quite costly to society if production of some valuable good or service is stopped (e.g. policeman on strike).

The threat of a strike or lockout is used as a negotiating tactic to secure a better contract for one of the sides. Although strikes and lockouts often receive considerable media attention (e.g. hockey or baseball players strikes), very few labor disputes end in a strike or lockout. This is likely due to the considerable cost to both sides from a strike or lockout. In addition, many states prohibit public sector employees from striking and instead force some sort of binding arbitration.

6.4 Rents and Unions

There is a connection between the product market structure and the effectiveness of unions in securing better labor contracts. As we discussed in a previous lecture, firms earn higher profits or rents in non-competitive product markets (e.g. a monopolist firm). Through collective bargaining, a unionized workforce may be able to extract some of these rents in the form of higher wages, benefits, or more jobs. In more competitive labor markets, competitive pressures may prevent a firm from allowing greater concessions to workers. In a highly competitive product market, where profits are near zero, a firm which pays its workers more than other firms may end up with negative profits and be forced to shut down. Therefore, we would expect
that unions would be more effective and have higher membership in less competitive industries.

An Example

One issue facing car manufacturers in the United States is the high cost of American labor (high wages and especially high pension and health care benefits) relative to foreign labor (e.g. relative to Japan, where national health insurance effectively subsidizes health care insurance costs for Japanese car manufacturers). In the period prior to the 1980s, where there was less foreign competition in the automobile industry, unions representing American auto workers were able to secure labor contracts with high wages and benefits. Now, with more competitive pressure on the automobile industry, there is greater resistance from American automobile firms to union demands. In particular, American automobile firms argue that they now need to reduce labor costs in order to compete with foreign manufacturers.

6.5 What Do Unions Want?

The objective of unions can vary according to their leadership and membership.

Some potential objectives:

1) Some unions may favor higher wages and benefits over higher employment.

2) Other unions may favor spreading out jobs and work hours among a large
number of employees in exchange for lower wages.

3) Still other unions may sacrifice health and pension benefits to avoid layoffs.

4) Other unions value compressing the wage structure to ensure that all workers, regardless of seniority or skill, receive similar wages.

### 6.6 Unions in the Public Sector

In the public sector, budgets for government services are set through some political process. Unlike private firms that compete with other firms, public sector firms (e.g., New York City) generally set their budgets with only limited competitive pressures.

In this environment, unions representing public sector workers (e.g., teachers, policeman, fireman) have two basic ways to increase the benefits to their workers:

1) *Increase Total Budget*

The union can lobby to increase the total government budget. If the share of labor costs in the total budget remains constant, an increase in the total budget would increase resources given to workers (through wages, benefits, or employment).

2) *Shift Resources to Labor*

The total budget remains the same, but the union can lobby to increase the share of the budget that is devoted to labor. This diverts resources from
other uses to pay labor costs. For example, a union representing teachers could lobby for resources to be shifted from school building maintenance or administrator pay to pay for higher teacher salaries or benefits.

Which of the two options public sector unions choose to pursue may have different implications. If it is thought that a particular public service is underfunded (e.g. not enough money for public schools), then increasing the total budget for this service would be beneficial. However, shifting resources from one public service input to another (e.g. from school building maintenance to teacher salaries) may be a mis-allocation of resources. It is unclear which inputs are most productive in improving the efficiency of the public service. For example, would it improve student learning more if we increased school building maintenance or teacher salaries?
7 Employment Contracting and Personnel Economics

7.1 Principal Agent Problems in Labor Markets

In the model presented so far, the labor market involves a simple transaction. Workers sell an hour of their time for an hourly wage. In this model, the market exchange is not very different from a farmer selling a bushel of wheat. A more realistic model of the labor market would recognize that the employment relationship is more complex. The primary factor influencing the complexity of labor market arrangements is the presence of principal-agent problems.

Principal-agent problems is a term used in economics for conflicts that arise between two economic actors. In the case of labor markets, the principal is the owner or owners of the firm. The agent are the workers that the firm employees. (Note: There could be multiple principal-agent problems within a firm as managers are agents for owners and the managers are also the principal for the workers they supervise.)

The root of the principal-agent problem is that the principal and agent have different objectives. The firm (principal) wants the worker (agent) to work as hard as possible at her job, but workers prefer lower effort.

There are four basic characteristics of principal agent problems in a labor market.
1) **Workers can decide to provide various amounts of effort on their jobs**

This assumption implies that the worker is not selling the firm a homogeneous good (a labor hour). A worker who is not working hard is often referred to as *shirking*.

2) **Effort is not costless for workers**

Workers do not like effort and would prefer low effort. If effort were costless, then there would be no principal-agent problem, as all workers would provide the maximum effort.

3) **A worker’s effort is not perfectly observed by the firm**

This assumption often arises because there is some *monitoring cost* to observing exactly how much effort a worker is providing. For example, it would be prohibitively costly for the firm’s owners to stand next to the worker all day long in order to assess how hard the employee is working.

Even if there are no monitoring costs, there may be other factors that affect a worker’s observed output which make it difficult for employers to measure an individual worker’s effort. If workers work in teams, separating one worker’s contribution from another worker’s is often not possible. If the effort of teams of workers cannot be separated, workers would have the incentive to *free ride* on the contributions of others by reducing their effort.

4) **Workers are not full claimants on their effort**
This assumption is really the definition of an agent. The worker receives pay from the firm’s owners, but does not necessarily receive all of the firm’s profits (or losses) that derive from the worker’s effort. A worker who is a full claimant on her effort would be self-employed, by definition. Various employment contracts have the goal of making workers nearly full claimants on their effort. Another way to express this is that many employment contracts try to align as closely as possible the objectives of the principal and agent by making.

7.2 Piece Rates

A piece rate employment contract bases a worker’s pay on their output or “piece”. In this system, workers are not paid for their time (either by the hour or a monthly salary). An example is paying a textile worker based on how many shirts she sews together.

7.2.1 A Model of Piece Rates

A piece rate wage takes the following form

$$w = \alpha + \beta q,$$

where $w$ is the worker’s pay and $q$ is the worker’s output. $\alpha$ and $\beta$ are parameters that determine how much the worker gets paid for each unit of output. Note that $\alpha$ indicates how much the worker gets if she produces
nothing \((q = 0)\). Here \(w\) is not a wage per hour, but the total pay the worker receives.

Main question: What is the optimal \(\alpha\) and \(\beta\) which will maximize the firm’s profits?

To answer this, we need to make a few more modeling assumptions.

Output

The output of worker depends on the worker’s choice of effort \((e)\).

\[
q = e
\]

Worker’s Problem

There is some cost to effort, \(c(e)\), with \(c'(e) > 0\). The utility function for the worker is simple. Utility is increasing in wages and decreasing in effort. The worker chooses the level of effort to maximize her utility. The worker’s utility maximization problem is

\[
\max_e w - c(e)
\]

\[
\max_e \alpha + \beta q - c(e).
\]

Substitute for \(q\):
The first order condition for this problem is

\[ \beta = c'(e). \]

\( \beta \) is the marginal benefit of additional effort. \( c'(e) \) is the marginal cost of additional effort. Note that \( \alpha \) doesn’t matter because the worker gets \( \alpha \) even with no output or effort.

**Firm’s Problem**

The firm chooses \( \alpha \) and \( \beta \) to maximize profits. The constraint on the firm’s problem is that these values of \( \alpha \) and \( \beta \) must be such that the worker will choose to work. The total benefit of working (\( w \)) for the worker must exceed the total cost (\( c(e) \)). The constraint is

\[ w \geq c(e) \]

Or

\[ \alpha + \beta e \geq c(e) \]

If \( \alpha + \beta e < c(e) \), then the worker quits and does not work at all.
The firm's maximization problem is

$$\max_{\alpha,\beta} q - w \ s.t. \ \alpha + \beta e \geq c(e),$$

where the price of output is normalized to 1, so total revenue is just $q$.

It is optimal for the firm to pay no more than the cost of effort. Therefore, the constraint becomes

$$\alpha + \beta e = c(e).$$

Substituting the constraint into the objective function, the maximization problem is then

$$\max_{\alpha,\beta} e - c(e).$$

The first order conditions are based on how the level of effort responds to changes in the wage parameters, $\alpha$ and $\beta$.

1) From the worker’s problem, we saw that the worker’s effort decision does not depend on $\alpha$

$$\frac{\partial e}{\partial \alpha} = 0$$

2) For $\beta$, 

\[
\frac{\partial e}{\partial \beta} - c'(e) \frac{\partial e}{\partial \beta} = 0.
\]

Re-arrange

\[
\frac{\partial e}{\partial \beta} [1 - c'(e)] = 0.
\]

The optimal \( \beta \) is for \( c'(e) = 1 \). From the worker’s problem, we know that \( c'(e) = \beta \). Thus the optimal \( \beta \) is \( \beta = 1 \).

To solve for \( \alpha \), we substitute the optimal \( \beta \) into the constraint: \( \alpha + \beta e = c(e) \).

Substituting this becomes

\[
\alpha + e = c(e).
\]

The optimal level of \( \alpha \) is \( c(e) - e \).

### 7.2.2 What Does this Mean?

The solution to this problem is for the firm to provide the worker the full benefit of her effort and output. By tying compensation directly to output, the firm (principal) has made the worker (agent) the full claimant on her output. To see this note that with \( \beta = 1 \) and \( \alpha = c(e) - e \), the worker gets paid
The wage is

\[ w = c(e) - e + q. \]

since \( q = e \).

With the optimal \( \beta \) and \( \alpha \), the worker is just indifferent between working and not. The total benefit of working is \( w \) which is set equal to the total cost of working \( c(e) \).

One way to interpret this payment scheme is that the firm is renting the means of production to the worker for the price of \( -\alpha \). The firm then allows the worker to collect all revenue from selling the product.

Some employment relationships are actually structured in exactly this way. Taxicab drivers often have an arrangement with their taxicab firms to rent the cab from the taxicab company. All the collected fares are kept by the taxi cab driver. The alternative contract is for the taxicab company to pay its drivers an hourly wage and then require the drivers to return all fares to them at the end of their shift.

The piece rate employment scheme where drivers rent their cabs and keep their fares is preferred for two reasons.
1) By allowing the drivers to keep all fares (making them full claimants on their effort), the taxicab company ensures that drivers work as hard as possible rather than snoozing on the side of the road.

2) Without allowing the drivers to keep all fares, the drivers would have an incentive to hide some fares and negotiate a separate, “off the meter,” deal with passengers. This is a type of monitoring problem.

7.2.3 Problems with Piece Rates

Piece rates are not used for all employment contracts. Some reasons why piece rates may not work:

1) *It is difficult to measure output*

   See discussion of monitoring costs above. How do we measure the output of managers, for example?

2) *Risk aversion*

   Making the worker the full claimant on her effort exposes the worker to risk. Any factor that reduces output leads immediately to losses in wages for the worker (e.g. a snowstorm stops all taxi cab traffic). To the extent that workers are risk averse, they may require higher wages to compensate them for these risks. Firms, because they are larger, may be better able to deal with risk than individual workers.

3) *Quantity/Quality Tradeoff*
Piece rate wages encourage workers to focus on maximizing the current quantity of their output. They may sacrifice investments in maintaining the quality of their capital equipment (e.g. maintenance of machinery) in favor of increasing the quantity of output. For example, taxicab drivers may drive their cabs very hard and neglect maintenance. This is a problem to the extent that quality is more difficult to measure than quantity. Otherwise the firm could attach piece rate type wages to quality as well. In general, there is a problem of providing workers incentives for one aspect of production if there are multiple production tasks.

### 7.3 Bonuses and Profit Sharing

Bonus and profit sharing share a similar structure with piece rates. The intention of these compensation schemes is to provide incentives for worker effort by basing pay on some measure of performance.

Bonuses are given to workers who meet either objective standards (e.g. sales targets) or subjective evaluations (e.g. a supervisor’s evaluation). For example, a bonus scheme could make the wage a function of sales:

\[ w = \alpha + \beta sales \]

Profit sharing ties compensation directly to the profits of the firm (\(\pi\)):

\[ w = \alpha + \beta \pi \]
This type of compensation scheme is found mainly among top managers of companies. One way to accomplish a profit sharing arrangement is for the firm to pay part of a worker’s compensation in the form of ownership shares of the firm (e.g. stocks or the option to purchase stocks, “stock options”).

The problems with these compensation schemes are similar to those with piece rates. First, it may be difficult to measure a worker’s entire performance based on a few objective or subjective measures. This is especially true when workers work in teams. Second, basing compensation strictly on profits or sales can cause workers to neglect other tasks. Some profit sharing plans where compensation is tied to a firm’s stock price may encourage managers to focus their effort on increasing the stock price and neglect the long term productivity of the firm.

### 7.4 Tournaments

Another compensation scheme designed to increase worker effort are tournaments. Some firms base compensation in the same way that sports tournaments are set-up. Workers start at entry level positions. If the workers perform relatively better than other entry level workers, they are promoted to a higher paying position. This competition continues until the worker reaches the top position in the firm. In principle, this is the same way a basketball tournament is structured. In the playoffs, basketball teams play a game and the winner moves on to play another game until only one team remains.
Notice two crucial differences between this type of compensation scheme and a piece rate, bonus, or profit sharing scheme.

1) In the tournament type of compensation system, *absolute* productivity or performance does not matter. For the worker to be promoted, she only needs to be better than her competition, her co-workers. Only *relative* performance matters.

2) Unlike the other compensation systems, in the tournament scheme the reward for higher effort now may not arrive until later when the individual is promoted.

Tournaments have been used as one explanation for the highly unbalanced compensation structure within firms. The highest paid employees within a firm (e.g. CEOs) are often paid many, many times more than entry level employees. The very high compensation for top managers may not reflect their productivity, but instead may exist to motivate lower level employees. The potential high pay for promotion in effect produces a tournament “prize,” which encourages higher levels of effort among employees competing for this prize.

The potential pitfalls of this type of compensation is that it discourages co-workers to collaborate, even if collaboration produces higher output. In fact, because tournaments value only relative performance, there are incentives to sabotage the performance of fellow employees. In these cases, the competition engendered by tournaments may be counter-productive. There-
Therefore, tournament schemes are often used by firms where employees act independently, such as independent salesman responsible for separate sales territories.

### 7.5 Deferred Compensation

Another compensation scheme changes how young workers (entry level) are paid relative to older workers (supervisors or managers). In the deferred compensation system, younger workers are paid less than their full marginal revenue product. This compensation is deferred until later. Older workers are paid more than their marginal revenue product.

The rationale for this scheme is that younger workers will increase their effort to avoid being fired. If a younger worker provides low effort, she may be fired and lose her future deferred compensation.

The problem with this type of compensation contract is that the firm has an incentive to fire older workers before they can collect their deferred compensation. If young workers know this, then they will not increase their effort. This problem can be circumvented with strict rules about firings and layoffs, which prevent the firm from arbitrarily eliminating expensive older workers. In addition, if a firm wants to maintain a reputation as a good employer in order to attract high quality workers, they will avoid reneging on deferred compensation agreements.
7.6 Efficiency Wages

When firms pay their own workers more than a worker could receive at other firms, these wages are called efficiency wages. The rationale for efficiency wages is that a firm wants to make itself the most desirable employer. Workers want to provide a high level of effort because they do not want to be fired from their current, more desirable, firm. If every other firm provided the same wage as a worker’s current firm, the costs of being fired are low and the worker may provide low effort.

An Efficiency Wage Model

The worker can either provide an effort of $e = 0$ (low effort) or $e = 1$ (high effort). The cost of high effort is $c$. The cost of low effort is 0. If the worker chooses effort of $e = 0$, she is fired with probability $p$. With probability $1 - p$ she is not fired. If the worker chooses $e = 1$, she is never fired. $p$ is assumed to be less than 1. This implies that the firm cannot perfectly monitor the worker’s effort and fire all the low effort employees.

All other firms offer the worker a wage of $w$. This is the worker’s outside option. If she is fired, she receives $w$ from an outside firm.

We want to show that it is optimal for the worker’s firm to provide an efficiency wage of $w^* > w$.

Worker’s Problem
If the worker chooses high effort, she receives

\[ u(e = 1) = w^* - c. \]

If the worker chooses low effort, she receives

\[ u(e = 0) = pw + (1 - p)w^*. \]

The worker will choose high effort if

\[ w^* - c \geq pw + (1 - p)w^*. \]

Simplifying,

\[ w^* \geq w + \frac{c}{p}. \]

**Firm’s Problem**

Assume the firm sets the efficiency wage at the minimum level to induce high effort. The firm sets the wage at

\[ w^* = w + \frac{c}{p}. \]

Is the firm receiving higher profits with this efficiency wage or with a lower wage \( w \)?
Assume that if the worker provides high effort, the firm produces $q^*$. If the worker provides low effort, the firm produces $q < q^*$.

Profits for the firm if the firm pays efficiency wages are

$$\pi^* = q^* - w^*.$$  

Profits for the firm if the firm pays $w$ are

$$\pi = q - w.$$  

The firm pays the efficiency wage if

$$q^* - w^* > q - w.$$  

Or

$$q^* - q > w - w^*.$$  

This implies that as long as the benefit in output due to higher worker effort ($q^* - q$) is greater than the cost of higher wages ($w - w^*$), the firm will pay the efficiency wage.
8 Compensating Differentials

The theory of compensating differentials states that workers are paid to compensate them for non-wage (non-pecuniary) aspects of jobs. Undesirable jobs, such as jobs with high risk of injury (e.g. policeman), or jobs with poor working conditions (e.g. coal miner), must offer higher wages in order to get people to work in these jobs. On the other hand, jobs with desirable characteristics, (e.g. good location, more benefits, or nice office furniture) can offer lower wages and still attract workers.

8.1 A Model of Compensating Differentials

To think about a compensating differentials model, re-write our labor supply utility function to include a $Z$ variable. $Z$ represents the non-pecuniary characteristics of a job. The utility function is then $U(c, l, Z)$. $Z$ is a desirable characteristic. Higher levels of $Z$ increase the worker’s utility

$$\frac{\partial U(c, l, Z)}{\partial Z} > 0$$

Suppose there are two firms. Firm A offers a wage $w_A$ and Firm B offers a wage of $w_B$. Both firms offer a package of non-pecuniary benefits: $Z_A$ and $Z_B$.

We make the following assumptions:

1) Firm A offers a higher level of non-pecuniary benefits: $Z_A > Z_B$. 
2) The wage offers are the same: \( w_A = w_B \).

Under these assumption, all workers will receive higher utility from working at Firm A. Therefore, every worker will choose to work at Firm A.

In order for Firm B to compete and attract workers, it must compensate the workers for the higher utility workers receive by working at Firm A. Suppose Firm B cannot adjust its level of non-pecuniary benefits (e.g., the firm is located in a bad location and cannot move). But Firm B can adjust its wage offer. In order to make workers just indifferent between Firm A and Firm B, Firm B must increase its wage offer.

To calculate the amount Firm B needs to increase its wages, first we need to calculate the indirect utility workers receive from a firm as a function of wages and non-pecuniary benefits. Let \( V(Z_A, w_A) \) be the amount of utility workers receive from working at Firm A. Workers receive \( V(Z_B, w_B) \) from Firm B. The difference in utility is then

\[
V(Z_A, w_A) - V(Z_B, w_B) > 0.
\]

Firm B calculates that it needs to raise wages to \( \bar{w}_B > w_B \) in order to equalize utility:

\[
V(Z_A, w_A) = V(Z_B, \bar{w}_B) = 0.
\]

Note that the wage offer Firm B offers is now greater than the wage offer
at Firm A: \( \bar{w}_B > w_A \). The difference in the wage offers, \( \bar{w}_B - w_A \), is the compensating differential for the difference in non-pecuniary characteristics, \( Z_A - Z_B \).

### 8.2 Using Compensating Differentials

Compensating differentials help to explain differences in wages across jobs and firms. Another use of compensating differentials is to enable economists to “price” characteristics that have no readily available prices. We could ask people how much they value characteristics of jobs. For example, we could survey people and ask them how much they would pay for their firm to move to a more desirable location. But it may be the case that this information would not be as accurate as using the actual observed behavior of people in the market.

consider two examples.

1) **What is the value of a human life?**

Two firms differ in the risk of death on the job. An example would be two coal mines and one of the mines has a higher risk of fatal accidents. At Firm A, the risk of death is \( p_A \). In one year, a worker has \( p_A \) probability of dying. The riskier mine, Firm B, has a risk of death of \( p_B > p_A \). Firm A pays an annual wage of \( w_A \). Firm B pays a wage of \( w_B > W_A \).

If we think that these wage differentials are compensation for the higher risk of death in Firm B, then we can use these differentials to calculate a
worker’s value of her life. Workers are willing to trade \( w_B - w_A \) dollars for a \( p_B - p_A \) higher probability of death.

Some Numbers

Assume that the probability of death is 0.001 greater in Firm B.

\[
p_B = p_A + 0.001
\]

The annual salary differential is

\[
w_B = w_A + 6,600
\]

Assume each firm employs 1,000 workers. The difference in the probability of death between the two firms implies that in a given year, one more worker will die in Firm B than in Firm A. Each worker in Firm B requires $6,600 in additional wages each year to compensate them for this risk. If there are 1,000 workers, Firm B is essentially “buying” one human life for 1,000 x $6,600 or 6.6 million dollars.

2) What is the value of school quality?

This is a non-labor market example. Some researchers have used differences in house prices to examine how much parents value differences in school quality.

Suppose here are two houses on opposite sides of the street: House A
and House B. The street forms a boundary between two school districts. The children from House A attend School A and the children from House B attend School B. The two schools have different levels of quality, measured by the average test scores of the students who attend the schools. The test scores for School A are $T_A$ and are greater than the test scores in School B: $T_A > T_B$. The price of House A is $P_A$ is higher than the price of House B: $P_A > P_B$. A compensating differentials model would indicate that we can use the difference in the prices of the houses, $P_A - P_B$, to measure how much parents value the difference in test scores, $T_A - T_B$.

### 8.3 Problems with Compensating Differentials

1) **People must know the actual differences.**

The compensating differentials model assumes that the decision makers know what the actual differences in characteristics are. The miners must know what the difference in the risk of death is between the two mines. The home owners must know the difference in test scores between the two schools. If decision makers (workers, homeowners) are acting on inaccurate information, then the compensating differential we observe in wages and home prices are meaningless.

2) **Must control for all other differences.**

Each of these examples is extremely simple. In reality, using compen-
sating differentials requires *controlling* for all other characteristics of job or
good. Ideally, the compensating differentials model only applies to situations
where the job or goods being compared are exactly the same, except for the
one characteristic we are interested in. For example, mines could differ in the
risk of death and in other factor such as life insurance benefits. The riskier
mine could offer its workers more life insurance benefits than the other mine.
The two houses could be very different from each other—one could have 2,000
square feet and the other only 1,000 square feet. If we do not take these
other factors into account in some way, the results from a compensating
differentials analysis are biased.
9 Human Capital

9.1 Human Capital Overview

Human capital is capital embodied in people. Like physical capital (e.g. machines, tools), human capital also makes labor more productive. In general, individuals with more human capital receive higher wages. The level of human capital in the population is a major reason why wages differ in the population.

The major sources of human capital are formal schooling, on-the-job training, and experience. However, there are many other types of human capital investments. Investments in health can be considered investments in human capital as healthy people are more productive. Another major source of human capital are the abilities and talents individuals are born with. The time and resources that parents spend caring for and raising their children can also be considered investments in their children’s human capital.

The level of human capital investments in the population varies widely. We will initially focus on schooling because it is a large source of human capital and is relatively easy to measure. Today, about 25 percent of the United States population (25 and older) have earned at least a college degree. Less than 20 percent of the adult population has not finished high school. Why individuals make different schooling decisions and the implications of these choices for the labor market are the topics of this section.
9.2 Human Capital and Productivity

To bridge the labor demand model we discussed earlier with this section on human capital, let’s assume the productivity of a worker is a function of human capital. Human capital can come from many different sources, but for simplicity, we can summarize the level of an individual’s human capital by the variable $S$. Write the marginal product of each labor hour as a function of human capital: $MP_h(S)$.

Because wages depend on the labor productivity, wages are now a function of human capital:

$$w(S) = p \cdot MP_h(S).$$

If we assume that more human capital makes workers more productive,

$$\frac{\partial MP_h(S)}{\partial S} > 0,$$

wages are therefore increasing in human capital:

$$\frac{\partial w(S)}{\partial S} > 0$$

9.3 A Model of Human Capital Investments

Consider a simple model of the decision to attend college. A recent high school graduate has two choices: she can start work right away or attend
college. If she attends college, after 4 years, she graduates from college and works with a college degree.

Assume that if she works without a college degree, she earns $w_H$. If she works with a college degree, she earns $w_C$. A reasonable assumption is that a college graduate earns more than a high school graduate: $w_C > w_H$. This assumption can be motivated based on the assumption that college makes workers more productive (college provides more human capital) or because college signals other forms of unobserved human capital (a signalling model, discussed below).

If attending college is costless, every person would choose to attend college. However there are some costs to attending college. One of the costs of obtaining a college degree is that, while the college student is in school, she cannot work (or at least not work as much if not attending college). There is an opportunity cost of college attendance from foregone earnings while in college.

9.3.1 Present Value Calculations

In order to model the costs of foregone earnings, we need to consider how individuals value present utility versus future utility. We use a concept called present value. It is a fair assumption that present utility is valued more than future utility. Said another way, future utility is discounted relative to present utility. The rate at which future utility is discounted is called the discount rate.
Discount rate

The discount rate is given by $\delta$: $0 \leq \delta \leq 1$. Higher $\delta$ indicates a higher value placed on future utility relative to present utility. If $\delta = 1$, future utility has the same value as present utility. If $\delta = 0$, future utility is completely discounted, and the individual only cares about present utility.

Assume there are two periods. $t = 1$ is today and $t = 2$ is tomorrow. If an individual receives $U_2$ in period 2 (tomorrow), the present value of this utility in period 1 (today) is

$$PV = \delta U_2$$

This indicates that $\delta U_2$ units of utility received tomorrow provides the individual less utility than receiving $U_2$ units of utility today.

$$PV = \delta U_2 \leq U_2$$

Interest Rates and Present Value

Present value calculations can be used to calculate the future value of money. In this case, the discount rate is a function of the interest rate: $\delta = \frac{1}{1+r}$. The present value of $\$100$ wage received in period 2 is

$$PV = \frac{\$100}{1+r}$$
An example: if the interest rate is 5 percent \((r = 0.05)\). The present value of $100 received in period 2 is

\[
PV = \frac{100}{1 + 0.05} \approx 95.2
\]

One justification for discounting the future in this case is that money can be invested today and earn interest collected tomorrow. If we could have $100 today, put it in a bank, and earn 5 percent interest, we could have $105 tomorrow. Therefore, having the $100 in hand today is more valuable than having the $100 tomorrow.

The present value calculation indicates that if we were given $95.2 today, we could save this money and at 5 percent interest, we would receive about $4.80 in interest. By saving the $95.2 today, we would receive approximately $100 tomorrow.

**Discounting More Than One Period**

To discount more than one period in the future, we use the following equation

\[
PV = U_1 + \delta U_2 + \delta^2 U_3 + \cdots + \delta^{T-1} U_T,
\]

where the periods are \(t = 1, \ldots, T\) and \(T\) is the last period. Notice that the farther away the future period is, the more it is discounted.
9.3.2 Foregone Earnings and College Attendance

Using the present value framework, we can write the present value of working after high school and not attending college. We assume a high school graduate earns $w_H$ every year she works. The high school graduate works every year until retirement in period $T$. Period $t = 1$ is the first year after high school graduation. The present value of earnings if an individual chooses not to attend college is

$$ PV_H = w_H + \delta w_H + \delta^2 w_H + \cdots + \delta^{T-1} w_H. $$

Or using $\sum$ notation, we can write this as

$$ PV_H = \sum_{t=1}^{T} \delta^{t-1} w_H. $$

Assume a college graduate also works every year after college graduation at a wage of $w_C$. A college graduate spends the first four years in college and earns no wages during these four years. The present value of earnings if the individual chooses to attend college is

$$ PV_C = 0 + \delta 0 + \delta^2 0 + \delta^3 0 + \delta^4 w_C + \delta^5 w_C + \cdots + \delta^{T-1} w_C. $$

Or using $\sum$ notation, we can write this as
These expressions provide the discounted lifetime earnings from the two choices. A high school graduate decides to attend college if the discounted lifetime earnings from attending college are greater than the discounted lifetime earnings of not attending college: $PV_C > PV_H$.

9.4 A More Detailed Look at Human Capital Investments

As the model now stands, there is still no reason why some people attend college and others don’t. If everyone has the same discount rate and earns the same wage from either choice, then the present value of the two choices is the same for all individuals. Either everyone should be attending college or no one should.

We can add several features to the model to make the model more realistic and provide some reasons for differences in the behavior observed in the population.

1) Direct Costs of Schooling

In the present model, the only cost to attending college is the opportunity cost of foregone labor market wages. There are also direct costs of attending
college, such as the cost of tuition. Assume the dollar value of these costs are $D$ ($D \geq 0$) per year. If $D$ does not vary over the four years of college attendance, the present value of college can be re-written as

$$PV_C = -\sum_{t=1}^{4} \delta^{t-1} D + \sum_{t=5}^{T} \delta^{t-1} w_C.$$ 

2) Ability and Wages

The model thus far assumes that all individual earn the same college wage (if they attend college) and the same high school wage (if they do not attend college). However, we know that there is substantial variation in wages even among the population of high school and college graduates. The wages individuals receive may be a function of other sources of human capital besides the human capital received from college. We can call these other sources of human capital ability. The source of these abilities may be from genetic endowments, from family or social environments, or from pre-college schooling (e.g. high school quality). Use $A$ as the scalar (one dimensional) ability endowment. Individuals with higher levels of $A$ have more ability. We can re-write the high school and college wage as a function of $A$: $w_H(A)$ and $w_C(A)$, where

$$\frac{\partial w_H(A)}{\partial A} \geq 0,$$

and
\[ \frac{\partial w_C(A)}{\partial A} \geq 0. \]

The present value of high school and college is now a function of an individual’s ability:

\[ PV_H(A) = \sum_{t=1}^{T} \delta^{t-1} w_H(A), \]

and

\[ PV_C(A) = -\sum_{t=1}^{4} \delta^{t-1} D + \sum_{t=5}^{T} \delta^{t-1} w_C(A). \]

Schooling decisions may vary because some people have higher abilities and this affects the wages they receive.

Consider two possibilities:

i) Ability and the human capital obtained from a college degrees are substitutes. At a high ability level, \( A_{\text{high}} \),

\[ PV_H(A_{\text{high}}) > PV_C(A_{\text{high}}) \]

In this case, high ability individuals receive sufficiently high wages without a college degree that the additional benefit from a college degree is not greater than the costs.
ii) Ability and college human capital are *complements*. At a high ability level, $A_{high}$,

$$PV_C(A_{high}) > PV_H(A_{high})$$

In this case, high ability individuals receive even higher wages with a college degree than without a degree. The combination of a college degree and high ability generates a large benefit to attending college relative to the cost.

3) *Tastes for Schooling*

In addition to variation in the abilities people have, individuals may also have different tastes for schooling. Some people may enjoy more (or dislike less) attending school more than others. To capture this idea, we need to depart from the idea that individuals only consider the present value of their lifetime earnings. Similar to the compensating differentials models, assume there is $Z$ variable that proxies for an individual’s taste for schooling. Higher values of $Z$ indicate a greater taste for schooling. In this model, individuals have a utility function which weighs the pecuniary aspects of schooling (wages and direct costs of schooling) with the non-pecuniary aspects given by $Z$. Write these utility functions as

$$U_H(Z, PV_H),$$
and

\[ U_C(Z, PV_C). \]

An individual chooses to attend college if

\[ U_C(Z, PV_C) > U_H(Z, PV_H). \]

Consider two cases:

i) If an individual has a high enough preference for schooling, she may attend college even if the present value of college earnings are lower than present value of earnings from working immediately.

ii) On the other hand, an individual with a high enough dislike for schooling may not attend college even if the pecuniary rewards are higher for college attendance.

4) Credit Constraints

Another important factor which may influence the decision to attend college is whether an individual can borrow to pay for the costs of college. In the model presented so far, there are direct costs to schooling because of tuition and the opportunity costs of not working while in school. The model implicitly assumes an individual can pay for these costs. However, as we know, many people borrow against future income in order to pay for college.
A potentially important source of heterogeneity in the population is that some individuals may be more credit constrained than others. Wealthy individuals (e.g. college students with wealthy parents) may have more money available to pay for college. Less wealthy individuals may have to borrow from banks or the government (student loans). These differences in wealth essentially impose higher credit constraints on some individuals than on others. It could be the case that an individual would receive a higher utility from attending college than not, but because they are credit constrained, they cannot afford the costs of college. Many government policies in the United States (e.g. financial aid for low income students, subsidies for public universities) are designed to alleviate credit constraints.

5) Uncertainty about Future Returns

Finally, it is important to consider that the wages paid to high school and college graduates may change over time. This is important for the decision to attend college to the extent individuals cannot predict with perfect certainty what the wage rates will be in the future. There is uncertainty about future returns to school. We can add a term to reflect this uncertainty to our college and high school wages:

\[ w_C(A, t) = w_C(A) + \varepsilon_C(t) \]
The wage college graduates receive $w_C(A, t)$ now varies by $t$. It is composed of two parts. $w_C(A)$ is the part of the wage college graduates of ability $A$ receive in all periods. This is the non-stochastic part of the wage. $\varepsilon_C(t)$ is the stochastic part of the wage. This term can be negative or positive in different time periods. If in some period $t$, the wage for college graduates is particularly high (e.g. because of high demand for college educated labor), the $\varepsilon_C(t)$ will be positive and large.

The same interpretation can be made for the high school wage $w_H(A, t)$. It is also composed of a non-stochastic part ($w_H(A)$) and a stochastic part $\varepsilon_H(t)$.

Because the wages are uncertain for human capital investments, there is some risk to investment in human capital. This risk can be thought of in the same way as the risk involved in investments in any other asset (e.g. the risk that the price of a stock in a firm might change in the future). Here the return to the college degree asset is the wage college graduates receive relative to the wage non-college graduates receive. To the extent that some people are more risk averse than others, the risk averse people may prefer not to invest in a college degree. The risk averse people want to avoid the risk that the return on this asset falls in the future.
An additional source of uncertainty is that an individual who attends college may not finish. This dropout risk adds additional risk to the investment in college human capital.

**Summary: Why Does Schooling Vary?**

The choice of attending college may vary in the population for several reasons:

1) Some people discount the future more than others. Individuals who value the present relatively more than the future would choose to work right away rather than waiting until after college graduation.

2) Some people have higher abilities than others and this affects the relative returns to college human capital.

3) Some people have a higher taste for schooling than others.

4) Some people are credit constrained and cannot afford to attend college.

5) The returns to college are uncertain and some individuals are more risk averse than others. These risk averse individuals choose not to invest in this risky asset (the college degree).

### 9.5 Life Cycle Human Capital Investment

Human capital investments take place throughout the life cycle, from birth to death. But, there is generally a distinct life cycle pattern where most
of the investment happens at the beginning of life and less human capital investment occurs toward the end of life. One strong reason for this pattern is that the earlier a human capital investment is made, the more periods there are for an individual to accumulate the return on the investment. An individual who graduates from college at age 22 has many more years to earn the higher wages associated with a college degree than an individual who graduates from college at age 40.

On the other hand, there are at least three main reasons why workers continue to make human capital investments throughout their life cycle.

1) **Human capital depreciates.**

   Just as a physical capital depreciates over time (e.g. tools become dull), human capital may also depreciate. Many students in this course have at least partially forgotten their high school math skills. Their accumulated stock of math human capital has depreciated. These students need to make new investments in math human capital in order to replace this depreciated portion of their human capital.

2) **Returns to human capital change over time.**

   Consider a secretary trained prior to the advent of personal computers in the 1980s. When the secretary was young in the 1970s, the return to investments in computer skills were low. She therefore decided not invest in computer human capital. In the 1980s, as personal computer technology
was introduced into the economy, the return to these investments increased. With the new higher return to investments in computer human capital, the administrative assistant now chooses to make these investments. The reason she needs to make these investments later in life is that she could not predict when she was young that the returns to computer human capital would rise. In this case, technological change creates uncertainty in the returns to human capital investments. This technological change affects the returns to human capital and causes workers to make new investments in human capital to update their skills.

3) **People update their preferences.**

When individuals are young they may have different preferences for education and occupations than when they get older. An individual when young may have had a strong preference for engineering and made investments in engineering human capital to work in an engineering occupation. As the individual got older, her preferences changed and she now has a strong preference for teaching. Because of this change in preferences, which were not perfectly predicted in youth, she must now make new investments in teaching human capital later in life.

### 9.6 Human Capital Production Functions

Human capital can be thought as an output which is produced using some combination of inputs. Like the production functions we discussed earlier, the
human capital production function describes the technology of how different combinations of inputs create the output of human capital:

\[ h = f(\text{purchased inputs, time, prior human capital}) \]

Higher levels of purchased inputs, time, or prior human capital are expected to increase the production of human capital \( h \).

Consider the output of labor economics human capital from this course. A student’s inputs into the production of labor economics human capital are her purchased inputs (tuition to pay the instructor’s salary, buy books), time (time spent in lecture and studying), and prior human capital. For this course, the student’s output of labor economics human capital may depend on how much economics and mathematical human capital the student has accumulated prior to the start of the course. Prior human capital may also depend on intelligence and other abilities a student is born with. All of these inputs would be expected to increase the production of labor economics human capital.

### 9.7 Complementarities in Human Capital Production

Complementarities in human capital investments occur when investments in past human capital increase the productivity of future human capital investments. Complementarities in human capital production may have important implications for policy interventions. Complementarities in human capital
production imply that early investments in childhood human capital development (e.g. pre-school programs) may be very cost effective relative to later interventions (e.g. job training for adults).

An Example: Cognitive Development

A large literature (much of it outside economics) documents that the later production of future human capital through schooling and job training is affected by the level of an individual’s cognitive ability (one imperfect measure would be IQ). These cognitive skills are partly determined by genetics and partly determined by early human capital investments (e.g. investments by parents in family environments and early schooling). The child development literature indicates that after about age 10, the level of cognitive ability is fixed and cannot be altered by later human capital investments. Prior to this age, investments in the human capital of children can influence cognitive development. To the extent that cognitive ability increases the productivity of later human capital investments, investment in cognitive human capital are *complementary* with later human capital investments. This feature of the human capital production function suggests the importance of investing in early cognitive development.
9.7.1 A Model of Human Capital Investments with Complements

There are two periods of decision making: $t = 1$ is childhood (all ages from birth to age 18), $t = 2$ is young adulthood (e.g. ages 18-24).

Investments in human capital can be made in either of the two periods. Call these investments in each period $I_1$ and $I_2$. This investment term can be thought of as the combination of the purchased inputs and time inputs.

Adult human capital (the human capital accumulated after period $t = 2$) is a function of all of the prior human capital investments:

$$h = g(I_1, I_2).$$

9.7.2 Two Extreme Cases

**Case 1: Perfect Complements**

If investments in period 1 and period 2 are perfect complements, adult human capital is given by this production function:

$$h = \min\{I_1, I_2\}.$$  

In the perfect complements case, early investments in $I_1$ must be matched equally by later investments in $I_2$. Any investment in period 2 larger than period 1 investments ($I_2 > I_1$) are wasted. Low investments in period 1
cannot be remedied with higher investments in period 2.

**Case 2: Perfect Substitutes**

In the case of perfect substitutes, adult human capital is produced as

$$h = aI_1 + (1 - a)I_2.$$ 

If this is the human capital production function, later investments in period 2 can substitute for low first period investments and make up for deficits in period 1 investments. At $a = 1/2$, there is no real difference between this two period childhood model and a one period model.

**9.7.3 Implications for the Timing of Interventions**

If the human capital production function exhibits perfect complementarity, an important implication is that inequalities in early human capital (e.g. cognitive development) at the point of late childhood and adulthood cannot be overcome with later human capital investments. This would suggest that subsidies for later human capital (e.g. secondary schooling, job training, and post-secondary financial aid) would be ineffective in reducing income inequality.

Even if there is only a limited degree of complementarity in human capital investments, early investments are preferred over later investments. As evidence for the presence of complementarities, there is a considerable amount
of research that finds that early interventions (e.g. subsidies for pre-school) have a large effect on adult outcomes (e.g. wages earned as adults).

An important issue to consider is why some parents do not make sufficient investments in early childhood human capital. Since there is a high return to these investments, less wealthy parents should be borrowing against future income to pay for these investments. The reason some parents do not borrow to pay for these investments is likely some form of a credit constraint. That is, parents cannot borrow against their children’s future income.

9.8 A Signalling Model of Human Capital Investments

Signalling models of human capital explain investments in observable human capital (e.g. formal schooling) as an attempt by workers to signal unobserved human capital to firms (e.g. the worker is intelligent, hard working, i.e. high ability). The difference between a signalling model of human capital and the prior models is that the observable human capital may not affect a worker’s productivity at all \( \frac{\partial MP_a(S)}{\partial S} = 0. \)

However, because firms believe that higher observed (unproductive) human capital signals unobserved (productive) human capital, workers with higher levels of observed human capital receive higher wages. The signalling model then provides a distinct explanation for the positive correlation in earnings and observed human capital.
9.8.1 Model Setup

There is one representative firm and two types of workers: Type 1 and Type 2.

Worker Types and Productivity

Type 1 workers have lower unobserved human capital than Type 2 workers. As a consequence, Type 1 workers have a lower productivity than Type 2 workers.

Type 1 worker’s marginal product is $q_1$.

Type 2 worker’s marginal product is $q_2 > q_1$.

Information

The proportion of workers of Type 1 is $p > 0$. Proportion of Type 2 is $1 - p$. The firm knows $p$ (the distribution of types in the population). But the firm does not know the type of any given worker before a wage offer is accepted. The firm learns the type of the worker only after the firm has paid the worker her wage.

In general, workers have no incentive to reveal their type. Type 1 workers always have the incentive to lie and claim they are high productivity Type 2 workers. In this environment, the firms cannot directly learn anything about a worker’s productivity.

Observed Human Capital (Schooling)
Observed human capital (schooling) is indexed \( s \). Where higher \( s \) indicates more schooling. The marginal cost (in $) to investments in schooling differs by type.

Type 1 marginal cost of schooling is \( c_1 > 0 \).

Type 2 marginal cost of schooling is \( c_2 > 0 \).

Type 1 has a higher marginal cost of schooling: \( c_1 > c_2 \).

One reasonable example of these assumptions would be that the smarter Type 2 workers have higher productivity and have a lower cost of obtaining schooling because school work is easier for them.

### 9.8.2 A Signalling Equilibrium

Workers of each type choose their investment in observed human capital (\( s \)). Firms choose a wage policy as a function of \( s \): \( w(s) \). A signalling equilibrium is where the beliefs of the firm/employer match the actual actions of the workers. Thereby the firm’s beliefs are confirmed by the workers’ choices, and there is no reason for the firm to change its beliefs.

Note that the signal in the model is schooling \( s \) and the signalling cost in the model is the cost of investing in schooling (\( c_1 \) and \( c_2 \)).

To find a signalling equilibrium in this model, we first choose a wage setting policy (\( w(s) \)). We then examine how workers choose their level of \( s \) in response to this wage policy.

Suppose the firm believes that there is some schooling level \( s^* \) such that all workers who choose \( s < s^* \) are the less productive Type 1 workers, and
all workers who choose \( s \geq s^* \) are the more productive Type 2 workers. The firm sets wages according to these beliefs (see graph):

\[
    w(s) = q_1 \quad \text{if} \quad s < s^*
\]

\[
    w(s) = q_2 \quad \text{if} \quad s \geq s^*
\]

How do the worker’s respond? Because of the structure of the wage setting policy and because there is some cost to schooling (signalling cost), a worker will either set \( s = 0 \) and receive \( w(s) = q_1 \) or set \( s = s^* \) and receive \( w(s) = q_2 \).

In order for the workers’ actions to confirm the firm’s beliefs, it must be the case that Type 1 workers choose \( s = 0 \) and Type 2 workers choose \( s = s^* \).

Type 1 chooses \( s = 0 \) over \( s = s^* \) if

\[
    q_1 - 0 > q_2 - c_1 s^*.
\]

Type 2 chooses \( s = s^* \) over \( s = 0 \) if

\[
    q_2 - c_2 s^* > q_1 - 0.
\]

Solve to find values of \( s^* \) which satisfy both inequalities:
\[
\frac{q_2 - q_1}{c_1} < s^* < \frac{q_2 - q_1}{c_2}
\]

Any values of \(s^*\) that satisfy this condition are signalling equilibria.

### 9.8.3 Some Comments

1) The model generates a positive correlation between observed human capital and wages, but observed human capital has no effect on productivity.

2) The key part of this model which allows firms to use observed human capital to signal worker type is that the cost of obtaining schooling is higher for the less productive workers. In general, a signalling equilibrium requires that the signalling cost must be negatively correlated with unobserved productivity.

3) There are multiple equilibria

There are a number of \(s^*\) that satisfy the condition above and generate a signalling equilibrium.

4) Different welfare implications for different values of \(s^*\)

Although for any \(s^*\), which satisfies the condition above, agents act rationally, the particular level of \(s^*\) has different welfare implications. Notice that an increase in \(s^*\) increases the cost of signalling for Type 2 workers but
does not affect Type 1 workers (they continue to choose $s = 0$). Therefore, the lowest $s^*$ creates the greatest total welfare among all $s^*$.

5) *Directly testing for type might be less costly*

Investments in human capital ($s$) appear to have no social return because they do not directly affect a worker’s productivity. But there is some benefit to signalling because these signals help solve an information problem (types are unobserved by firms). However, there may be less socially costly ways for firms to distinguish between workers and allocate workers correctly. Firms may be able to use some sort of test to provide information about a worker’s type.

**9.8.4 A No Signalling Regime**

It is possible that workers may prefer a no signalling regime in which there are no productive signals.

Consider a model where $s$ provides no signal of type. In this model, firms base their wage offers on the expectation of worker types. All workers receive the same wage:

$$w = pq_1 + (1 - p)q_2.$$  

Type 1 workers would prefer this no signalling wage since $p > 0$ and $q_2 > q_1$. Type 2 workers also may prefer the no signalling wage if
\[ q_2 - c_2 s^* < pq_1 + (1 - p)q_2 \]

Notice several comparative statistics about this relationship. The benefit to Type 2 workers of the no signalling wage relative to the signalling wage increases with

i) The smaller the difference in productivity between types \((q_1 - q_2)\).

ii) The higher the cost of signalling \((c_2)\).

iii) The higher the fraction of Type 2 workers in the economy \((1 - p)\).

iv) The higher the signalling point \(s^*\).

Given that both types of workers may prefer a no signalling wage, but firms in general would prefer to receive signals, there may be some scope for workers of different types to collude. Type 1 and Type 2 workers could both agree to choose \(s = 0\), thus eliminating the value of the signal and returning to the no signalling wage.

### 9.9 Non-Schooling Human Capital

Another major source of human capital is the human capital individuals obtain from work experience and post-schooling training. We can think of this type of human capital as rather heterogenous, as it is often specific
to firms, occupations, and industries. Non-schooling human capital can be informally obtained on-the-job through learning by doing. Or, it can be obtained more formally through training courses the worker or the firm pay for.

In general, evidence for the importance of non-schooling human capital is more difficult to obtain given that much of it is not measured in major surveys. This is in contrast to schooling human capital (e.g. college measured by a college degree), which is relatively easier to measure.

To examine the importance of non-schooling human capital, economists use indirect evidence. In particular, we know that wages continue to increase with age even after formal schooling has been completed. We can write wages $W$ as a function of age:

$$W = f(age)$$

This *age-earnings* profile can be represented by a graph with age on the horizontal axis and wages or earnings on the vertical axis. Wages are increasing in age:

$$\frac{\partial W}{\partial age} = \frac{\partial f(age)}{\partial age} > 0.$$ 

There is some evidence that this pattern is due to investments in non-schooling human capital individuals continue to make as they get older. The
next section examines investments by firms in their non-schooling human capital through on-the-job training.

9.10 On-the-Job Training

Training that is provided by firms is called on-the-job training. Firms invest in the human capital of their workers because it makes the workers more productive. This increase in labor productivity can in some situations increase the firm’s profits.

9.10.1 Two Types of On-the-Job Training

Firms can make two types of investments in a worker’s human capital.

1) General Human Capital

   General human capital is defined as human capital that is productive in more than one firm. For a welder employed by the automaker GM (her firm), general human capital would be welding skills the welder could use in many firms and industries.

2) Firm Specific Human Capital

   Firm specific human capital is only productive in a given firm. For a welder employed by GM, firm specific human capital would be the knowledge the welder has in welding together the unique parts for GM cars. This human capital is not productive in another firm.
We know that firms will generally provide firm specific human capital to their workers. However, we do not know whether they will also provide general training. The extent to which firms provide general training to their workers is potentially one of the main determinants of the stock of human capital in an economy. If firms do not provide general training and workers cannot finance these investments themselves (e.g. because of credit constraints), there may be a rationale for government subsidized training.

### 9.10.2 Will Firms Provide General Training to their Workers?

The answer depends on the structure of the labor market.

1) *Perfectly Competitive Labor Market*

If the labor market is competitive, firms will not invest in general skills because workers will leave after the training. A firm that pays for the general training of their workers makes these workers more productive with all firms. If the firm does not increase the worker’s wages to reflect this higher productivity, the worker will leave the current firm and work in another firm. Even if the firm raises the worker’s wages to reflect the increased productivity, the cost to the firm of the higher wages completely offsets the benefit to the firm of higher productivity. The firm therefore has no incentive to invest in general human capital.

In contrast, a firm invests in firm specific skills because these skills are only productive with the current firm.
Workers may still invest in general skills on the job by paying for the training themselves through lower wages. As an example, apprentices typically are paid less than their marginal revenue product during the apprenticeship period.

2) Non-Competitive Labor Market

As we discussed earlier, in a non-competitive labor market there are frictions or imperfections in the labor market which impose costs on workers moving to another firm. This provides the firm an incentive to invest in the general skills of their workers since the firm can capture at least part of the surplus associated with the higher productivity of workers with their current firm relative to other firms.

To illustrate the point, take the extreme case of slavery. Slavery amounts to a complete labor market friction in which the firm owns the worker, and the worker cannot leave the employer. In this example, the firm has the incentive to make optimal investments in the worker’s human capital through general training. Since the worker cannot leave, the firm can capture all of the benefits from the investment in the worker’s general human capital.

9.10.3 Model of On-the-Job Training

There are two periods in the model and two agents: a worker and an employer (firm).

Period 1
During the first period, there is a joint decision by the worker and employer on the investment in general human capital. For simplicity, there is no firm specific human capital in the model.

General human capital is measured by $\tau \geq 0$.

Production in period 1 is normalized to 0.

The worker receives a first period wage of $W$.

**Period 2**

During the second period, the worker either stays with the employer and gets a second period wage $w(\tau)$ or quits the firm and receives her outside option from another firm ($v(\tau)$). Note that the second period wages and outside options are a function of $\tau$. It is reasonable to assume that

$$\frac{\partial w(\tau)}{\partial \tau} \geq 0$$

$$\frac{\partial v(\tau)}{\partial \tau} \geq 0$$

**Outside Option**

What the worker receives if she quits is key. The outside option depends on the market structure of the labor market:

i) Perfectly competitive labor market: worker gets $v(\tau)$. 
ii) Non-competitive labor market: worker gets \( v(\tau) - \Delta \), where \( \Delta > 0 \). \( \Delta \) represents the cost to the worker of the friction in the labor market. The presence of these frictions distinguishes a non-competitive labor market from a competitive labor market.

*Exogenous Separation*

With probability \( q \), the worker and firm receive an exogenous adverse shock, which causes them to separate (e.g. a recession causes the firm to shut down). \( q \) is a measure of expected turnover in the model, aside from voluntary quits by the worker.

*Optimal General Training Level*

The worker produces \( f(\tau) \). This function is increasing in the training level \( \tau \).

\[
f'(\tau) = \frac{\partial f(\tau)}{\partial \tau} \geq 0
\]

The cost of acquiring general human capital is \( c(\tau) \). \( c(\tau) \) is increasing in \( \tau \).

\[
c'(\tau) = \frac{\partial c(\tau)}{\partial \tau} \geq 0
\]

The optimal training level is \( \tau^* > 0 \). \( \tau^* \) is the solution to this first order condition:
\[ c'(\tau^*) = f'(\tau^*). \]

\( \tau^* \) is the optimal training level which sets the marginal cost of training equal to the marginal benefit.

**Perfectly Competitive Labor Market**

The distinguishing feature of perfect competition is that if the worker quits, she gets the outside offer of \( v(\tau) = f(\tau) \), i.e. \( \Delta = 0 \). That is, her outside option reflects exactly her productivity with the current firm.

Here firms do not invest in training even though the optimal training level is \( \tau^* > 0 \). Because firms cannot lower worker wages to pay for training, the worker immediately leaves for her outside option as soon as any training is received. Anticipating this, firms do not make any investments in general training.

However, workers can receive training if they choose to pay the full costs of training themselves through lower wages. Workers are the full residual claimant on their training investment because they can move freely between firms. The worker therefore choose the optimal training level \( \tau^* \). The first period wage is

\[ W = -c(\tau^*) \]
Non-Competitive Labor Market

Frictions in the labor market imply that general training has less value in an outside firm than in the current firm: \( v(\tau) < f(\tau) \). This means there is a surplus of \( f(\tau) - v(\tau) \) for the worker and firm to bargain over.

A useful way to summarize this is that the wage the worker receives in the second period is

\[
w(\tau) = v(\tau) + \beta[f(\tau) - v(\tau)],
\]

where \( \beta \) is the bargaining power of the worker. If \( \beta = 1 \), the worker gets all of the surplus. If \( \beta = 0 \), the firm gets all of the surplus.

As long as \( \beta < 1 \) and \( q < 1 \), a sufficient condition for positive training investment by the firm is \( f'(0) > v'(0) \). The basic intuition is that an investment in training (at the point of \( \tau = 0 \)) increases profits for the current firm more than it increases the outside option of the worker.

As we might expect, higher bargaining power of the worker or higher quit rates decrease the incentive for firms to invest in training.

An important point is that the profit maximizing investment in general training by the firm is generally less than the optimal training investment \( (\tau^*) \). This is because the firm is not the full residual claimant on the increased productivity from training investments. Two reasons for less than optimal training investments:
1) The worker receives part of the surplus back in wages, depending on the bargaining power parameter.

2) The risk of exogenous separation ($q > 0$) creates uncertainty that the firm could lose their investment in the worker’s general skills.

9.10.4 Sources of Labor Market Frictions

1) Search and Mobility Costs

See the discussion of search and mobility costs we discussed earlier. If these types of labor market frictions exist, the worker has an incentive to stay with the current firm and the firm has an incentive to invest in the general training of the worker.

2) Asymmetric Information I

Outside firms have less information about the quality and content of training a worker receives. Therefore, outside firms may not pay the full value of the training a worker receives. A third party credential or certification system could be designed to solve this problem.

3) Asymmetric Information II

The current firm learns about the imperfectly observed ability of workers. This information cannot be readily shared with other firms. An adverse selection (“lemons market”) problem occurs because a worker sends a signal
of low ability by being laid off by the firm (we expect low ability workers
to be fired or laid off first). This allows firms to keep and train high ability
workers and pay them lower wages than they would otherwise receive if all
firms had the same information.

4) *Complementarity between General and Firm Specific Skills*

If firm specific skills make general skills more productive, investments in
firm specific skills increase the return to general skills with the current firm
more than with outside firms.
10 Econometrics Review

10.1 Random Variables and Data

A random variable is a placeholder for an unknown experimental outcome. One example is flipping a coin. Let $X$ be the random variable for flipping a coin. Before we flip the coin and know the result of the experiment, we know that $X$ can either be heads or tails.

$$X = \{ \text{"heads"}, \text{"tails"} \}$$

Each $X$ outcome ("heads" and "tails") is a realization of the random variable $X$.

If the die is fair, the probability that $X$ is heads is 0.5. The probability that $X$ is tails is also 0.5. The following is the probability distribution for the random variable $X$.

$$pr(X = \text{"heads"}) = 0.5$$

$$pr(X = \text{"tails"}) = 0.5$$

This probability distribution tells us the probability of each and every outcome or realization for the random variable $X$.

A more relevant example is a survey of labor hours. Assume 5 individuals are surveyed. Each survey respondent is asked how many hours they worked
last week. Let $Y_i$ be the random variable for the hours of work individual $i$ reports. There are 5 random variables corresponding to each respondent’s hours of work:

$$Y_1 = 0, Y_2 = 40, Y_3 = 20, Y_4 = 47, Y_5 = 38$$

This collection of hours of work information is our data. Each individual’s report is a data observation. There are 5 observations in our data.

A non-random variable is a constant. This is simply a number which has no probability distribution (or a “degenerate” distribution).

### 10.2 Descriptive Statistics

Functions of data are called statistics. For this labor hours data, we can calculate these descriptive statistics (or sample statistics):

1) **Sample Median**

   This is the middle number of hours arranged from highest to lowest.

   $$med(Y) = 38$$

2) **Sample Mean or Sample Average**:
\[ \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{1}{5}(0 + 40 + 20 + 47 + 38) = \frac{145}{5} = 29, \]

where \( N \) is the number of observations in the sample.

The symbol used for the Sample Mean, \( \bar{Y} \), is said “Y bar”.

3) **Sample Variance**

Sample variance tells us something about the dispersion of data around the mean.

\[ S^2_Y = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \]

\[ S^2_Y = \frac{1}{5}[(0 - 29)^2 + (40 - 29)^2 + (20 - 29)^2 + (47 - 29)^2 + (38 - 29)^2] \approx 289.6 \]

Variance is never negative: \( S^2_Y \geq 0 \). The higher the variance, the more dispersion there is in the data.

4) **Sample Standard Deviation**

The sample standard deviation is the square root of the variance.

\[ S_Y = \sqrt{S^2_Y} = \sqrt{289.6} \approx 17. \]
10.3 Types of Variables

1) Discrete Variables

Discrete variables take on a countable number of values. For example, the variable for a coin flip is discrete because it can only take on two values: “heads” and “tails.”

2) Dummy Variables

Dummy variables are discrete variables which take on only one of two values: 0 or 1. The value of the dummy variable is intended to indicate some characteristic. For example, we can summarize the gender of the survey respondents by a dummy variable $D$. $D = 1$ if the person is male, and $D = 0$ if the person is female.

3) Continuous Variables

Continuous variables take on a potentially infinite number of values. Hours worked last week is a continuous variable. This variable can take any values from 0 to 168.

10.4 Populations and Samples

Samples are drawn from populations. In our hours of work survey data, the population is all adults in the United States. The sample of 5 individuals is drawn from this population of over 100 million people.
The major distinction between a sample and a population is that statistics can be calculated for a sample, but these same statistics are in general unknown for the population. We can calculate the mean hours worked for our sample of 5 survey respondents, but we can never know the mean hours worked in the entire population. The exception to this is experiments where the population probability distribution is known (e.g. the coin flip experiment). For these known experiments, we can calculate population statistics.

The reason we collect data on a sample of individuals rather than the population is that it is in general infeasible to collect information on the entire population. We hope that samples tells us something about the population. We use sample statistics to infer information about the population. The connection between the sample and population is therefore called inference.

10.5 Sampling Schemes

There are two main ways to randomly sample from a population:

1) Random Sample

A random sample collects information from people in a population chosen at random. A practical way to do this for our hours of work survey would be to pick random telephone numbers and survey each person who picks up the phone. The reason we use random samples rather than non-random samples is that a random sample of individuals better represents the population. If we were to only survey lawyers or only survey 30 year old males, we would
obtain biased statistics for the population we are interested in (all American adults).

2) *Stratified Sample*

Stratified sampling is random sampling for particular sub-groups or *strata*. The population is divided into sub-groups by characteristics such as race, gender, occupation, or residence. Within each sub-group a random sample is collected. The reason many surveys are collected as stratified samples is to ensure there are adequate numbers of people with a given set of characteristics in the sample. For example, a non-stratified random sample of a 1,000 people in the United States may only contain a handful of people from New York. If we want reliable statistics for New Yorkers, we should instead stratify our sampling on location and collect at least a 100 observations from New York.

10.6 Consequences of Random Sampling

Drawing a random sample means that the randomly collected observations have two properties:

i) Observations are *independent*.

Independent observations have no relationship to each other; they were collected randomly. We did not survey our five closest friends, but instead randomly chose people to survey.
ii) Observations are *identically distributed*.

If observations have an identical distribution, they are drawn from the same population. Each observation reflects the population it was drawn from and has the same population distribution. There is no fundamental difference between one randomly collected observation and another.

10.7 Expectations

The expectation of a random variable is essentially the population equivalent of the sample mean (or the sample mean is the *sample analog* of the population expectation). Except in the case of controlled experiments (e.g. the coin flip experiment), the expectation of a random variable cannot be known because it is a function of the unknown population distribution of a random variable.

For a discrete random variable $X$, the expectation of this random variable is

$$E(X) = \sum_x pr(X = x)x,$$

where $x$ (lower case) is one of the outcomes or realizations of the random variable $X$, $pr(X = x)$ is the probability distribution of $X$, and $\sum_x$ indicates that we are summing over all of the possible outcomes or realizations of $X$.

(Note: For continuous variables, expectations is defined using integrals because there are an infinite number of realizations of the random variable.)
For a continuous variable $X$, $E(X) = \int x f(x) dx$, where $f(x)$ is the continuous probability distribution function for $X$.

**Properties of the Expectations Operator**

For random variables $X$ and $Y$ and constants $a$ and $b$,

i) $E(b + aY) = b + aE(Y)$

ii) $E(aY + bX) = aE(Y) + bE(X)$

iii) $E(YX) \neq E(Y)E(X)$ (in general).

### 10.8 Relationship between Variables: Variance, Covariance, Correlation, and Independence

#### 10.8.1 Variance

The variance of a random variable is the population equivalent of the sample variance.

$$V(X) = E[(X - E(X))(X - E(X))] = \text{Var}(X).$$

Sometimes the notation $\text{Var}(X)$ is also used for variance.

**Properties of the Variance Operator**
For random variables $X$ and $Y$ and constants $a$ and $b$,

i) $V(b) = 0$. Variance of a constant (non-random variable) is zero.

ii) $V(b + aX) = a^2 V(X)$.

iii) $V(aX + bY) = a^2 V(X) + b^2 V(Y) + ab 2 \text{cov}(X,Y)$

### 10.8.2 Covariance

Covariance indicates the relationship between two variables:

$$\text{cov}(X,Y) = E[(X - E(X))(Y - E(Y))].$$

Notice that the covariance between the same variables is the variance: $\text{cov}(X,X) = V(X)$. The covariance operator therefore has the same properties as the variance operator.

**Interpretation of Covariance:**

i) If $\text{cov}(X,Y) = 0$, there is no relationship between $X$ and $Y$.

ii) If $\text{cov}(X,Y) > 0$, there is a positive relationship between $X$ and $Y$ (higher $X$ is associated with higher $Y$).

iii) If $\text{cov}(X,Y) < 0$, there is a negative relationship between $X$ and $Y$ (higher $X$ is associated with lower $Y$).
Properties of the Covariance Operator

For random variables, $X$ and $Y$, and constants $a$ and $b$.

i) $\text{cov}(X, X) = V(X, X)$

ii) $\text{cov}(X, Y) = \text{cov}(Y, X)$

iii) $\text{cov}(a, b) = 0$. Covariance of constants is zero.

iv) $\text{cov}(a, X) = 0$. Covariance of a constant and a random variable is zero.

10.8.3 Correlation

Correlation is a units free measure of a relationship between two variables. The correlation coefficient between $X$ and $Y$ is written $\text{corr}(X, Y)$.

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{V(X)^{1/2}V(Y)^{1/2}}$$

Correlation is bounded between $-1$ and $1$.

$$-1 \geq \text{corr}(X, Y) \geq 1,$$

Correlation $\text{corr}(X, Y)$ has the same sign and interpretation as covariance $\text{cov}(X, Y)$:

i) $\text{corr}(X, Y) = 0$ indicates no correlation between $X$ and $Y$.

ii) $\text{corr}(X, Y) = 1$ indicates a perfect positive correlation between $X$ and $Y$. 

iii) $corr(X, Y) = -1$ indicates a perfect negative correlation between $X$ and $Y$.

10.8.4 Independence

Independent random variables have no relationship with each other. The notation for independence is $\perp$. If $X$ and $Y$ are independent, we write $X \perp Y$. Note that $X \perp Y$ implies $Y \perp X$.

All independent variables have a covariance and correlation of 0.

$$X \perp Y \Rightarrow cov(X, Y) = 0$$

$$X \perp Y \Rightarrow corr(X, Y) = 0$$

However, variables with a covariance and correlation of 0 are NOT necessarily independent.

10.9 Conditional Expectations

Means, variances, and covariances can be calculated conditionally. $E[Y|X]$ is the expectation of $Y$ conditional on $X$ (mean of $Y$ conditional on $X$). Conditioning on a random variable transforms the conditioning random variable $X$ into a constant. The conditional mean is now a function of the conditioning variable $X$. 
An Example

$D$ is a dummy variable for male gender ($D = 1$ for men, and $D = 0$ for women) and $Y$ is hours worked last week. $E[Y|D = 1]$ is the expectation of hours worked for men. $E[Y|D = 0]$ is the expectation of hours worked for women.

We can write this conditional expectation as a function:

$$E[Y|D] = \beta_0 + \beta_1 D,$$

where $\beta_0$ and $\beta_1$ are population parameters.

The conditional expectations for each value of $D$ are

$$E[Y|D = 1] = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

and

$$E[Y|D = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

Independence and Conditional Expectations

It is important to note that if any two random variables $X$ and $Y$ are independent ($X \perp Y$), then $E[Y|X] = E[Y]$ and $E[X|Y] = E[X]$. The intuition is that if $X$ and $Y$ are not related to each other, then conditioning on the other will not affect their expectation.
In our example above, $D \perp Y$ implies that gender ($D$) has no relationship with $Y$. Hence, the conditional expectations of $Y$ for men and women are the same. If $D \perp Y$, on average, men and women work the same number of average hours $E[Y]$.

If $D \perp Y$, then

$$E[Y|D = 1] = E[Y],$$

and

$$E[Y|D = 0] = E[Y].$$

**Conditional Variances and Covariances**

Conditional variances and covariances have the same interpretation. For example, $V(Y|D = 1)$ is the variance of hours worked for men. $V(Y|D = 0)$ is the variance of hours worked for women.

**10.10 Sample Analog**

We can construct sample analogs for each of the population concepts defined above. The *sample analog* is the sample statistic calculated using a data sample of size $N$.

i) Mean, Average

Sample analog: $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$

ii) Variance


Sample analog: $S_Y^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$.

iii) Standard Deviation


Sample analog: $S_Y = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2}$.

iv) Covariance

Population: $cov(X,Y)$.  

Sample analog: $S_{XY} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})$.

v) Correlation

Population: $corr(X,Y)$.  

Sample analog: $\hat{corr}(X,Y) = \frac{S_{XY}}{S_X S_Y}$.

vi) Conditional Expectations


Sample analog: we calculate the sample mean of $Y$ for each value of $X$.

For our dummy variable example above, the sample analog of $E[Y|D = 1]$ is

$$
\bar{Y}_{male} = \frac{1}{N_{male}} \sum_{i=1}^{N_{male}} Y_i,
$$

where $\bar{Y}_{male}$ is the sample average of $Y$ for the males in our sample, and $N_{male}$ is the number of males in our sample.

The sample analog of $E[Y|D = 0]$ is

$$
\bar{Y}_{female} = \frac{1}{N_{female}} \sum_{i=1}^{N_{female}} Y_i,
$$

where $\bar{Y}_{female}$ is the sample average of $Y$ for the females in our sample, and $N_{female}$ is the number of females in our sample.

vii) Conditional Variance


Sample analog: we calculate the sample variance of $Y$ for each value of $X$.

For our dummy variable example above, the sample analog of $V[Y|D = 1]$ is

$$
S^2_{Y,male} = \frac{1}{N_{male}} \sum_{i=1}^{N_{male}} (Y_i - \bar{Y}_{male})^2.
$$
10.11 Relationship between Sample and Population Statistics

From our sample, we calculate sample statistics. We would like these statistics to be as close as possible to the unknown population statistics. Consider the difference between the population and sample mean.

Call the unknown population mean $\mu$. $\mu$ is NOT a random variable. Unless we have a survey the entire population, we cannot know the value of $\mu$.

Our sample mean is $\overline{Y}$. The sample mean is a function of our sample data. Each data observation is a random variable. Because the sample mean is a function of random variables, it is also a random variable. The sample mean is one estimator of the unknown population mean.

Is the sample mean a “good” estimator of the unknown population mean?

10.11.1 Bias

One criteria for deciding how “good” an estimator is to examine its bias.

\[
\text{bias} = E(\overline{Y}) - \mu.
\]

We would like this bias to be as close to 0 as possible.

What is the expectation of the sample mean?
Using the properties of expectations, we can write this as

\[ E(\bar{Y}) = \frac{1}{N} \sum_{i=1}^{N} E(Y_i) \]

Because of random sampling, each observation has the same distribution and the same population mean. For all observations \( i \), \( E(Y_i) = \mu \).

Substituting this,

\[ E(\bar{Y}) = \frac{1}{N} \sum_{i=1}^{N} \mu = \frac{1}{N} (\mu + \cdots + \mu) \]

Since \( \mu \) is a constant, we can re-write this as

\[ E(\bar{Y}) = \frac{1}{N} \ast N \ast \mu = \mu \]

This shows that the bias of the sample mean is 0 in a random sample. The sample mean is an unbiased estimator of the population mean.

### 10.11.2 Variance of the Estimator

Another desirable property of an estimator is low variance. High variance implies that for different samples we might have widely different sample mean
estimators. High variance therefore reduces the \textit{precision} of our estimator.

The variance of the sample mean estimator is

\[ V(\bar{Y}) = V\left(\frac{1}{N} \sum_{i=1}^{N} Y_i\right) \]

Using the variance operator properties,

\[ V(\bar{Y}) = \left(\frac{1}{N}\right)^2 \left[ V(Y_1) + V(Y_2) + \cdots + V(Y_N) + 2\text{cov}(Y_1, Y_2) \right. \]

\[ + 2\text{cov}(Y_1, Y_3) + \cdots + 2\text{cov}(Y_N, Y_{N-1}) \]

Because of random sampling, observations are independent. For any two observations \(i\) and \(j\), we know \(Y_i \perp Y_j\). Independence implies that all of the covariance terms are 0. Since the covariance terms are 0, we can re-write the variance of the sample mean as

\[ V(\bar{Y}) = \left(\frac{1}{N}\right)^2 \sum_{i=1}^{N} V(Y_i) \]

Because of random sampling, all of the observations \(i\) have an identical distribution. This implies that \(V(Y_i) = \sigma^2\) for all \(i\). \(\sigma^2\) is the unknown population variance. Like \(\mu\), \(\sigma^2\) is NOT a random variable.

Substituting,

\[ V(\bar{Y}) = \left(\frac{1}{N}\right)^2 N\sigma^2. \]
Simplifying,

\[ V(\bar{Y}) = \frac{1}{N} \sigma^2. \]

This expression indicates that the variance of our sample mean estimator is a function of the unknown population variance \( \sigma^2 \).

### 10.11.3 Estimating the Population Variance

Without an estimator for the unknown population variance, we cannot estimate the variance of the sample mean estimator. A sensible estimator for the population variance \( \sigma^2 \) is the sample variance of \( Y \), \( S_Y^2 \).

\[ S_Y^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \]

Substituting this into the expression for the variance of the sample mean (above), we obtain an estimator for the variance of the sample mean:

\[ \hat{V}(\bar{Y}) = \frac{1}{N} S_Y^2. \]

### 10.11.4 Standard Error for the Sample Mean

The square root of the estimated variance of the sample mean is called the standard error for the sample mean.
\[ SE(\bar{Y}) = \sqrt{V(\bar{Y})} = \frac{1}{\sqrt{N}} \sqrt{S_Y^2} \]

This statistic provides a measure of the precision of the sample mean estimator. Low standard error indicates a highly precise estimator.

Returning to our hours of work example

In our hours of work survey data there are \( N = 5 \) observations. The sample mean for our hours of work survey is \( \bar{Y} = 29 \). The sample variance of \( Y \) is \( S_Y^2 = 289.6 \), and \( \sqrt{S_Y^2} \approx 17 \).

The standard error of our sample mean estimator is then

\[ SE(\bar{Y}) = \frac{1}{\sqrt{5}} \times 17 \approx 7.60. \]

10.12 Inference for the Sample Mean

Our goal is to learn something about the population mean from a random sample. We know that because of random sampling, the sample mean is an unbiased estimator of the population mean. But the sample mean is a random variable. There is some probability that the sample mean could be very different from the unknown population mean.

We would like to have an idea of the probability distribution of the sample mean. If we know the distribution of the sample mean, we can calculate
the probability that the sample mean is very different from the unknown population mean. We cannot know exactly what the distribution of the sample mean estimator is. But, under some assumptions (embodied in a Central Limit Theorem), we can approximate this distribution.

10.12.1 Confidence Intervals

Using the approximation of the distribution of the sample mean, we can construct the following confidence interval:

With 95 percent probability, the unknown population mean \( \mu \) is within this interval:

\[
\bar{Y} - SE(\bar{Y}) \times 2 \leq \mu \leq \bar{Y} + SE(\bar{Y}) \times 2,
\]

where 2 is the critical value, which depends on the confidence level of the interval. The critical value for the 95 percent confidence level is 2. (Note: the critical value is not exactly 2, but we’ll just use 2 in this course.)

For our example, this confidence interval is

\[
29 - 7.60 \times 2 \leq \mu \leq 29 + 7.60 \times 2
\]

\[
13.8 \leq \mu \leq 44.2
\]
It is possible $\mu$ is not equal to the sample mean of 29. However, given the approximation of the distribution of the sample mean, we can say with 95 percent confidence that $\mu$ is in the interval between 13.8 and 44.2.

Comments on Confidence Intervals

1) Smaller confidence intervals indicate greater precision in the estimator.

2) A higher number of observations (higher $N$) reduces the size of the confidence interval.

3) A lower standard error ($SE(\bar{Y})$) reduces the size of the confidence interval.

4) If we want to be even more confident about the range of values the unknown $\mu$ could take on (e.g. raise the confidence level to 99 percent), then we need to increase the critical value and the confidence interval becomes wider. For example, the 99 percent confidence interval has a critical value of about 2.6.

10.12.2 Hypothesis Tests

Confidence intervals can be used to conduct tests of specific hypothesis. A hypothesis which is often tested is whether the unknown population mean is 0. If the value of 0 falls within the 95 percent confidence interval, then we fail to reject the hypothesis (at the 95 percent confidence level) that $\mu = 0$. In our example, 0 is outside the calculated 95 percent confidence interval. We
can therefore say that we reject the hypothesis that $\mu = 0$ at this confidence level.

10.13 Regression Analysis

10.13.1 Regression Model

The regression models we will study take this form:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

This is a *theoretical* model about how values of $Y$ relate to values of $X$ in the population. Notice that this regression model is *linear* in $X$. This regression model provides an equation for a line, a regression line.

$Y$ is the *dependent* variable. It is a random variable. The variable $Y$ depends on the value of the independent variable $X$.

$X$ is the *independent* or *explanatory* variable. It is also a random variable. The $X$ variable is also sometimes called the *regressor*.

$\beta_0$ and $\beta_1$ are the regression model *parameters*. $\beta_0$ is the intercept of the regression line. $\beta_1$ is the slope of the regression line.

Values of $\beta_1$ have the following interpretation:

1) If $\beta_1 > 1$, then $Y$ is increasing in $X$: $\frac{\partial Y}{\partial X} > 0$. 
2) If $\beta_1 < 1$, then $Y$ is decreasing in $X$: $\frac{\partial Y}{\partial X} < 0$.

3) If $\beta_1 = 0$, then $Y$ is not related to $X$: $\frac{\partial Y}{\partial X} = 0$.

$\varepsilon$ is sometimes thought of as a the error or residual component of a regression model.

The major distinction between $\varepsilon$ and the $X$ and $Y$ variables is that we observe the $X$ and $Y$ variables, but the $\varepsilon$ variable is unobserved. That is, we can survey individuals and ask them to tell us their $X$ and $Y$ values. In our hours of work survey, we could ask the survey respondents to tell us their hours of work (their $Y$) and tell us their gender (their $X$ variable, where $X = 1$ if male, and $X = 0$ if female). Whatever we do not observe in data we include in the $\varepsilon$ random variable.

Regression Model Assumptions

The regression model we will maintain has the following four main assumptions:

1) $\varepsilon$, $Y$, and $X$ are random variables. We do not know their population distributions (their population means, variances, covariances, etc.).

2) The relationship between $Y$ and $X$ is linear and given by this equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

3) $E[\varepsilon] = 0$. 
4) $\varepsilon \perp X$.

This last assumption is the most important. This assumption is often referred to as an *exogeneity* assumption. $X$ is assumed to be an *exogenous* variable in the regression model. If $X$ is not independent of $\varepsilon$, then $X$ is an *endogenous* variable.

Note an immediate implication of this assumption is that $\text{cov}(\varepsilon, X) = \text{corr}(\varepsilon, X) = 0$.

There are other important regression model assumptions, but for simplicity we will not discuss these.

*The Regression Model Holds for Everyone*

This regression model holds for all units in the population (e.g. individuals, firms, etc.). To emphasize this, let’s index each random variable by a subscript $i$. $i$ indicates a particular unit (e.g. an individual) from the population. Note that here $i$ does not indicate a data observation because this regression model is a theoretical model for the population.

The model written with subscripts $i$ is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Notice that the parameters are not indexed by $i$. One of the implicit assumptions of the regression model is that these parameters are the same for everyone.
For simplicity, we often simply drop the $i$ subscripts. But, unless stated otherwise, it is always the case that a given regression model holds for all individuals in the population.

### 10.13.2 A Prediction Interpretation of Regression Models

Consider another interpretation of regression models. Let’s assume our goal is to *predict* the $Y$ value for someone. If we don’t know anything about the person, the “best” predictor of the $Y$ for a given person is the average $Y$, $E[Y]$. However, if we know some more information about the person, say the individual’s $X$ value, we may be able to form a better prediction of the person’s $Y$ value IF there is a relationship between $X$ and $Y$.

With no information about $X$ values, our model for each individual $i$ is simply

$$Y_i = \beta_0 + \varepsilon_i$$

Given the assumptions of the regression model, $\beta_0$ is simply $\beta_0 = E[Y_i]$. Our predicted $Y_i$ is indicated $\tilde{Y}$. The predicted value of $Y$ in this case is $\tilde{Y} = E[Y]$.

In this interpretation of the regression model, $\varepsilon_i$ is the *prediction error*. For each individual indexed $i$, the prediction error is

$$\varepsilon_i = Y_i - \tilde{Y} = Y_i - E[Y_i] = Y_i - \beta_0 = \varepsilon_i$$
The assumption above that $E[\varepsilon] = 0$ (or, equivalently, $E[\varepsilon_i] = 0$) is an assumption that the average prediction error is zero. That is, $\bar{Y}$ is an unbiased predictor of $Y_i$.

If we know more information, an $X_i$ value for each individual $i$, our model is then

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Our “best” predictor uses this $X_i$ variable information to form a better predictor of an individual’s $Y_i$ value. The predictor is now the expectation of $Y$ conditional on $X$: $\bar{Y} = E[Y_i|X_i]$. According to the assumptions above, this conditional expectation has a specific (linear) form:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

$\varepsilon_i$ is again the prediction error:

$$\varepsilon_i = Y_i - \bar{Y} = Y_i - E[Y_i|X_i] = Y_i - \beta_0 + \beta_1 X_i = \varepsilon_i$$

The assumptions of the regression model again imply that this prediction error is mean zero ($E[\varepsilon_i] = 0$). However, the predictor using $X$ variables is potentially a better predictor given the additional information $X$ provides about individuals.

It is important to note that if there is no relationship between $X$ and $Y$,
then this model collapses back to the model without the \( X \). If \( X \) and \( Y \) are independent, then \( \beta_1 = 0 \). The model is then

\[
Y_i = \beta_0 + 0X_i + \varepsilon_i = \beta_0 + \varepsilon_i.
\]

**An Example: Predicting Wages**

Assume our goal is to predict a person’s wage \((W_i)\). If we know nothing about the individual, the best predictor of the individual’s wage is the average wage, \( E[W_i] \). However, we think there is a relationship between an individual’s human capital (measured by the years of schooling the individual has completed, \( S_i \)) and the individual’s wage. We think this relationship is given by this regression model:

\[
W_i = \beta_0 + \beta_1 S_i + \varepsilon_i
\]

If there is some relationship between \( S_i \) and \( W_i \), then we can use this information to form a better predictor of an individual’s wage. For example, an individual with a college degree (\( S_i = 16 \)) likely has a higher wage than an individual with no college degree (\( S_i = 12 \)). This positive relationship between human capital and wages would be indicated by a positive \( \beta_1 \). Our regression model uses schooling information to form a better predictor of an individual’s wage. We will discuss this problem in more detail below.
10.13.3 Estimating Regression Model Parameters

Given a sample of data (a $Y$ value and $X$ value for each person or data observation), we can estimate the $\beta_0$ and $\beta_1$ parameters. In many circumstances, the optimal method for estimating regression model parameters is called *Ordinary Least Squares* or OLS. Call the estimators for the population parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. The OLS equations for these estimators are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$  

$S_{XY}$ is the *sample covariance* of the $X$ and $Y$ variables in the data sample:

$$S_{XY} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})$$

$S_X^2$ is the *sample variance* of the $X$ variable.

$$S_X^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

$\bar{X}$ is the sample mean of the $X$ variable, and $\bar{Y}$ is the sample mean of the $Y$ variable.

Notice that the equation for $\hat{\beta}_1$ implies that this estimator reflects the
covariance in the $X$ and $Y$ data.

If $X$ and $Y$ have positive sample covariance, $\hat{\beta}_1 > 0$.

If $X$ and $Y$ have negative sample covariance, $\hat{\beta}_1 < 0$.

If the sample covariance between $X$ and $Y$ is zero, $\hat{\beta}_1 = 0$.

10.13.4 Inference

Because the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of the random variables $Y$ and $X$, they are also random variables and have some unknown distribution. Just as we did for the sample mean estimator, we can examine the bias and variance of our estimators. It turns out, that if the assumptions of the regression model hold, the OLS estimators are unbiased:

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$

*Estimating the Standard Error of the OLS Estimators*

Under the assumptions of our regression model and some additional assumptions (which we will not discuss), the variance of the estimator for $\beta_1$ is
\[ V(\hat{\beta}_1) = \frac{\sigma^2}{S_X^2}. \]

As with the variance of the sample mean estimator, we need to estimate the unknown population variance \( \sigma^2 \). Using the assumptions of the regression model, an estimate of \( \sigma^2 \), called \( \hat{\sigma}^2 \), can be found as

\[ \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left[ Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right]^2 \]

With this estimate for \( \sigma^2 \), an estimate of the variance of the estimator \( \hat{\beta}_1 \) is

\[ V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_X^2}. \]

The standard error of \( \hat{\beta}_1 \) is then

\[ SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)}. \]

Like the standard error for the sample mean, we want as low a standard error as possible. Lower standard error for \( \hat{\beta}_1 \) indicates a more precise estimate of \( \beta_1 \).

The \( SE(\hat{\beta}_1) \) decreases with
i) Lower variation in $Y$ (lower $\hat{\sigma}^2$).

ii) Higher variation in $X$ (higher $S_X^2$).

iii) More observations (higher $N$).

The variance and standard error for $\hat{\beta}_0$ have a similar expression and interpretation.

Confidence Intervals and Hypothesis Tests for the OLS Estimators

Confidence intervals and hypothesis tests for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be constructed in the same way as we constructed them for the sample mean estimator.

The 95 percent confidence interval for $\beta_1$ is

$$\hat{\beta}_1 - SE(\hat{\beta}_1) \times 2 \leq \beta_1 \leq \hat{\beta}_1 + SE(\hat{\beta}_1) \times 2.$$ 

10.13.5 Multivariate Regression Models

Our regression model thus far has one $X$ variable and is a univariate model. Regression models with more than one $X$ variable are called multivariate regression models.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

The estimators for the multivariate regression parameters ($\beta_0$, $\beta_1$, $\beta_2$) have different OLS equations than in the univariate model. However, the
basic interpretation of the parameters is similar. In addition, the inference, standard errors, and confidence interval concepts carry over to the multivariate framework.

The advantage of multivariate regression models is that more explanatory $X$ variables can be used to predict the $Y$ variable. For example, if $Y$ is hours of work last week, $X_1$ could be a dummy variable for gender, and $X_2$ could be a variable for the wage rate of the individual. With this model, we could estimate whether men and women have different labor supply responses to changes in wage rates. The applications we discuss next will present more material on how to interpret regression analysis in particular contexts.
11 Topics in Applied Labor Economics: Estimating the “Return” to Schooling

11.1 What is the “Return” to Schooling?

For a variety of reasons, economists have been interested in estimating the “return” to schooling. What exactly this “return” to schooling represents is an open question. One interpretation is that the return to schooling is the causal effect of forcing a random person to complete additional schooling. The assumption is that this individual would earn more in the labor market (higher wages) with the additional schooling than without the additional schooling. The increase in labor market earnings is assumed to reflect the increase in the labor productivity of the individual due to the higher level of human capital she now has.

Let’s look at the return to college, where for simplicity we’ll define college as 16 years of schooling and non-college as 12 years. The expected return to college can be defined as

\[ E[W_i | S_i = 16] - E[W_i | S_i = 12], \]

where \( S_i \) is the years of schooling variable, \( W_i \) is labor market earnings (wage), and \( i \) indexes individuals.

We can re-write this in terms of a regression model as
\[ W_i = \beta_0 + \beta_1 S_i + \varepsilon_i. \]

\( \varepsilon_i \) is the stand-in for all of the other unmeasured factors that affect wages.  
\( \varepsilon_i \) is simply the residual difference in wages, net of the “effect” of schooling:

\[ \varepsilon_i = W_i - (\beta_0 + \beta_1 S_i) \]

Using the assumptions of the regression model, the expectation of the wage for individuals with 12 and 16 years of schooling are

\[ E[W_i|S_i = 16] = \beta_0 + \beta_1 16 \]

\[ E[W_i|S_i = 12] = \beta_0 + \beta_1 12 \]

The difference is

\[ E[W_i|S_i = 16] - E[W_i|S_i = 12] = \beta_1 4. \]

\( \beta_1 \) indicates how much the wage increases with a one unit (one year) increase in schooling.  To see this explicitly, re-write the regression model to indicate that the wage is a function of \( S_i \) and \( \epsilon_i \).
\[ W_i(S_i, \varepsilon_i) = \beta_0 + \beta_1 S_i + \varepsilon_i \]

The partial derivative of the wage function with respect to schooling is

\[ \frac{\partial W_i(S_i, \varepsilon_i)}{\partial S_i} = \beta_1 \]

### 11.2 Percent Change

The previous regression model has wage levels as the dependent variable and \( \beta_1 \) indicates the change in the level of wages from a change in schooling. However, we are often more interested in the percent change “effect” schooling has on wages. To calculate this, we transform the wage using the log function. Re-write the regression model as

\[ \ln W_i(S_i, \varepsilon_i) = \beta_0 + \beta_1 S_i + \varepsilon_i. \]

Now calculate the partial derivative,

\[ \frac{\partial \ln W_i(S_i, \varepsilon_i)}{\partial S_i} = \frac{\partial W_i(S_i, \varepsilon_i)}{\partial S_i} \frac{1}{W_i(S_i, \varepsilon_i)} = \beta_1 \]

\( \beta_1 \times 100 \) indicates the percent change in the wage with a one unit increase in schooling. For example, if \( \beta_1 = 0.07 \), an individual who increase her
schooling 1 more year will have 7 percent higher wages. An individual who increases her schooling 2 more years will have 14 percent higher wages, and so on.

Note: the $\beta_1$ and $\beta_0$ in the model with wage levels is not the same as the $\beta_1$ and $\beta_0$ in the model with log wages.

11.3 OLS Estimation

The OLS estimators for this regression model are

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(\ln W_i - \ln \bar{W})(S_i - \bar{S})}{\frac{1}{N}\sum_{i=1}^{N}(S_i - \bar{S})^2}$$

$$\hat{\beta}_0 = \ln \bar{W} - \hat{\beta}_1 \bar{S}$$

I estimated this model using data from the March 2003 Current Population Survey. The number of observations is $N = 87,585$.

You can obtain this data at this website:

https://beta.ipums.org/cps/

The regression model estimates are (standard errors are in parentheses).

$$\hat{\beta}_1 = 0.09216 (0.00069)$$
\[ \hat{\beta}_0 = 1.488 \ (0.00967) \]

This coefficient estimate indicates that each year of schooling increases wages by about 9.2 percent.

11.4 Inference

Using an approximation of the unknown distribution of \( \hat{\beta}_1 \), we can construct a 95 percent confidence interval for the \( \beta_1 \) parameter. With 95 percent probability, the unknown \( \beta_1 \) population parameter lies in this interval:

\[
0.09216 - 2 \times 0.00069 \leq \beta_1 \leq 0.09216 + 2 \times 0.00069
\]

\[
0.0908 \leq \beta_1 \leq 0.0935
\]

This is a fairly tight confidence interval.

We can also conduct hypothesis tests. Since 0 is outside the 95 percent confidence interval, we can reject, with 95 percent probability, the hypothesis that \( \beta_1 = 0 \), and the return to schooling is zero. However, we cannot reject the hypothesis, with 95 percent probability, that \( \beta_1 = 0.091 \), for example.
11.5 Self-Selection Bias

There is a strong reason to suspect that this OLS estimate of $\beta_1$ is biased. An OLS regression estimator may be biased if one or more the regression model assumptions does not really hold.

Recall that one of the assumptions of the regression model is that the explanatory variable ($X$ variable) and the error component ($\varepsilon$) are independent. In this case, this assumption holds if the number of years of schooling an individual obtains ($S_i$) is independent of the $\varepsilon_i$ variable. To understand whether this is in fact true, we need to think about what the $\varepsilon_i$ variable reflects.

In our regression model, we partitioned the factors that explain an individual’s wage ($W_i$) into two factors: years of schooling, which we observe in our data, and all other factors represented by $\varepsilon_i$. As we discussed above, it is likely that the wage an individual receives is affected by more than just the level of formal schooling an individual has obtained. For example, other sources of human capital, which we called ability, may also affect the wage an individuals receives.

Let’s re-write our regression model to include a random variable for an individual’s level of ability ($A_i$).

$$\ln W_i(S_i, A_i, \eta_i) = \beta_0 + \beta_1 S_i + \alpha A_i + \eta_i,$$

where $\alpha$ is a population parameter which indicates the relationship be-
between ability and wages (e.g. if ability increases an individual’s wage, then $\alpha > 0$). $\eta_i$ is another random variable (error component) which reflects everything else that affects wages.

Since we do not observe $A_i$ or $\eta_i$ in our data set, without loss of generality, we simply called these terms $\varepsilon_i$. That is, $\varepsilon_i$ in our original regression model is

$$\varepsilon_i = \alpha A_i + \eta_i$$

In the human capital lecture above, we also discussed why an individual’s level of ability may affect their human capital investment (e.g. their choice to attend college). Let’s write another regression model to indicate that there is a relationship between schooling and ability.

$$S_i = \gamma_0 + \gamma_1 A_i + \omega_i$$

This model indicates that the level of an individual’s schooling ($S_i$) is related to the level of the individual’s ability ($A_i$) and to other factors represented by $\omega_i$. $\omega_i$ reflects all of the other factors that affect an individual’s schooling choice (e.g. individual $i$’s discount rate, taste for schooling, degree of credit constraint, etc.).

If $\gamma_1$ is not zero (i.e. there is some relationship between ability and schooling), then there is some correlation between schooling for an individual and
that individual’s ability. Therefore, $\text{cov}(S_i, A_i) \neq 0$. Since $\varepsilon_i$ is also a function of ability, $\text{cov}(S_i, \varepsilon_i) \neq 0$. If this is the case, the regression model assumption that the explanatory variable and the error component are independent does not hold.

Without showing this formally, this violation of the regression model assumptions implies that the OLS estimator we used above provides a biased estimate of the regression model parameters $\beta_0$ and $\beta_1$. In particular, we think that higher ability individuals are more likely to attend school and earn higher wages (regardless of schooling level). If this is true, then the OLS estimator for the return to schooling has an upward bias: $\hat{\beta}_1 > \beta_1$.

*Five Ways to Express This Bias*

i) *Self-Selection Bias*

This bias is often called *self-selection bias* because the source of the bias is from the fact that individual’s self-select into schooling, i.e. schooling is not randomly assigned.

ii) *Sample Selection Bias*

This bias can also be thought of as a *sample selection bias*. The root cause of the bias is that we do not have a random sample. The sample of individuals who attend college is not a random or representative sample of the entire population. The sample of individuals who attend college is a selected or *choice based* sample.
iii) *Omitted Variable Bias*

Another name for this bias is *omitted variable bias*. The bias in this case stems from the fact that we cannot observe all aspects of an individual’s human capital. $A_i$ is therefore omitted from the regression model we can estimate. If we were able to observe $A_i$ completely, then we could include it in our regression model, and the omitted variable bias would be eliminated.

iv) *Endogenous Regressors*

Still another term for this bias is to say that $X$ is an *endogenous regressor*. Recall that if $X$ is independent of $\varepsilon$, then $X$ is an exogenous regressor. In this case, if $\text{cov}(X, \varepsilon) \neq 0$, then $X$ is an endogenous regressor.

v) *Correlation is Not Necessarily Causation*

Finally, we can also think of this bias as reflecting the fact that observed correlations in random variables do not necessarily reflect causation. For almost any data set, including our CPS data, the sample correlation wages and schooling is positive: $S_{SW} > 0$. However the self-selection bias suggests that this correlation may not be entirely due to the *causal effect* of schooling on wages.

11.6 Difference in College and High School Wages

Let’s return to examining the difference between someone with 12 years of schooling (high school graduate) and someone with 16 years of schooling (col-
lege graduate). The expectation of their wage conditional on their schooling is

\[ E[W_i | S_i = 16] = \beta_0 + \beta_1 16 + E[\varepsilon_i | S_i = 16] \]

\[ E[W_i | S_i = 12] = \beta_0 + \beta_1 12 + E[\varepsilon_i | S_i = 12] \]

Assume that individuals self-select into college based on ability. Assume that because of this self-selection, the average ability for those who attend college is \( E[A_i | S_i = 16] = A_{col} \). The average ability for individuals who do not attend college is \( E[A_i | S_i = 12] = A_{nc} \). Given these assumptions, we can write:

\[ E[\varepsilon_i | S_i = 16] = \alpha A_{col} + E[\eta_i | S_i = 16] \]

\[ E[\varepsilon_i | S_i = 12] = \alpha A_{nc} + E[\eta_i | S_i = 12] \]

Assume that \( \eta \) is independent of \( S_i \). Therefore,

\[ E[\eta_i | S_i = 16] = E[\eta_i | S_i = 12] = 0. \]

Substituting back into the conditional expectation functions for wages,
Taking the difference, 
\[ E[W_i|S_i = 16] - E[W_i|S_i = 12] = \beta_1 4 + \alpha (A_{col} - A_{nc}) \]

The expected difference in wages is the left-hand-side of this equation. This difference consists of two parts:

i) the difference in schooling multiplied by the return to that difference in schooling (\(\beta_1 4\)),

and

ii) the difference in average ability between the college educated and non-college educated populations multiplied by the return to that ability difference (\(\alpha (A_{col} - A_{nc})\)).

If we believe that \(\alpha > 0\) (higher ability individuals receive higher wages), and \((A_{col} - A_{nc}) > 0\) (college educated individuals have on average more ability than non-college educated individuals), then the difference in average
wages is an upwardly biased indicator for the actual return to schooling \((\beta_14)\).

Self-selection essentially implies that the correlation between schooling and wages we observe cannot be interpreted as entirely due to causation.

### 11.7 A Treatment Effects Interpretation

Another way to interpret this self-selection bias is to think of a college education as a *treatment*. In one sense, we would like to know the causal effect of taking a person randomly from the population and forcing this person to take the treatment, i.e. graduate from college and push her \(S_i\) from \(S_i = 12\) to \(S_i = 16\). How much would this randomly chosen person’s wage increase? If the assumptions hold, the regression model answers this question. The wage for this person would increase by \(\beta_14\). \(\beta_1\) indicates the *treatment effect* on wages of increasing an individual’s schooling.

If people randomly choose their schooling, then there is no difference in the average ability level of the people who have \(S_i = 12\) and those who have \(S_i = 16\). Randomization of schooling implies \(S_i\) is independent of \(\varepsilon_i\) and the assumptions of the regression model hold. Randomization implies that

\[
E[\varepsilon_i|S_i = 12] = E[\varepsilon_i|S_i = 16]
\]

However, in reality, because people do not randomly choose to enter college, but instead self-select into college based on factors (e.g. ability) which also affect their wages, these statements break down.
11.8 Possible Solutions to the Self-Selection Bias

Possible Solution 1: Conduct a Controlled Experiment

The most immediate possible solution to this problem is to conduct a controlled experiment. In the experiment, we force some randomly chosen individuals to attend college (the treatment group). Another group of people (the control group) is forced not to attend college. An unbiased estimator of the return to schooling can be constructed simply by comparing the average wage for the two groups. This type of experiment is no different in principle from a randomized drug trial in which a randomly selected treatment group receives the drug being tested, and a control group receives a placebo.

More feasible experiments may be to randomly offer some people free college tuition (the treatment group) and not offer the aid to others (the control group).

Although these types of experiments may offer excellent information on the return to schooling, these types of experiments are relatively rare in the social sciences. For many interesting and important questions, a randomized experiment is unethical, impractical, or too expensive.

Possible Solution 2: Measure Omitted Variables

Another possible solution is to try to measure the omitted variable. In this case, we might use an IQ test or SAT score to measure an individual’s level
of $A_i$. Assume we have an $IQ_i$ measure for each individual. Our regression model is then

$$\ln W_i(S_i, IQ_i, \varepsilon_i) = \beta_0 + \beta_1 S_i + \beta_2 IQ_i + \varepsilon_i.$$  

Including measures of the omitted variable in this way may reduce the self-selection bias, but it is unlikely that it will eliminate the bias since most measures, like IQ and SAT scores, are only partial measures of the factors that affect an individual’s wages.

**Possible Solution 3: Instrument or Natural Experiment**

Another way to approach this problem is to look for variables which are i) correlated with an individual’s schooling level, but ii) not correlated with wages except through schooling (i.e. variables not correlated with $\varepsilon_i$). These types of variables are called *instrumental variables*.

Consider an example. One instrumental variable approach has been to use the proximity of an individual to college or university as an instrumental variable. The idea is that living closer to a college or university lowers the cost of attending college. It can also be argued that the proximity to a college or university should not affect an individual’s wages (except through the schooling).

Why does this help solve the self-selection problem? The idea behind using instrumental variable is that these variables essentially form a *natural experiment*. For whatever reason, some individuals, and not others, live
close to colleges and universities and the cost of schooling is relatively low for this group. This variation in proximity to college approximates an actual experiment in which the cost of schooling is randomly changed for some individuals and not for others. As in a controlled experiment, we can use this source of *exogenous variation* to solve the self-selection bias in the OLS estimation.

It is important to understand the limitations of this approach. In general, the natural experiment or instrumental variable only affects the behavior of particular group of individuals. In the example of college proximity, the group affected is the group on the margin between deciding whether to attend or not attend college. For most individuals, the choice of whether to attend college or not would not be affected by the proximity of a college or university. Because the natural experiment only uses the behavior of a particular group, the return to schooling using this variation may not tell us very much about the return to schooling for most other individuals.

*Possible Solution 4: Use Identical Twins*

Another interesting possible solution to the self-selection bias is to compare identical twins. The idea is that twins share a common genetic and family background. If this common genetic and family background captures all of the factors that affect wages (our $A_i$ term), then the difference in wages for two twins with different schooling levels provides an unbiased estimator of the return to schooling.
Index each pair of twins in the population by $i$. Label the wage of twin 1, $W_{i1}$, and the wage of twin 2, $W_{i2}$. The schooling level of each twin pair is given by $S_{i1}$ and $S_{i2}$. For each pair of twins, we difference their wages and schooling levels:

$$W_{i1} - W_{i2} = \beta_1 (S_{i1} - S_{i2}) + \varepsilon_{i1} - \varepsilon_{i2}$$

The assumption underlying the use of twins is that each pair of twins has the same $\varepsilon_i$: $\varepsilon_{i1} = \varepsilon_{i2}$.

Our estimator estimates the expectation of the difference in wages for all twins:

$$E[W_{i1} - W_{i2}] = \beta_1 E[(S_{i1} - S_{i2})]$$

A potential problem with this methodology is that it relies on at least some twins having different levels of schooling ($S_{i1} \neq S_{i2}$ for at least some $i$). Why did otherwise identical twins choose different levels of schooling? The answer may be that identical twins, although they may have the same genetics, are not exactly the same. These differences between twins may be correlated with schooling choices and wages. If this is true, we are right back to the original self-selection problem.

*Possible Solution 5: Model Selection into Schooling*

The final approach to the self-selection problem uses economic theory to
model how people choose their human capital investments. If the model is correct, we can find an unbiased estimator of the return to schooling. The estimation based on these models is often called \textit{structural estimation}. The disadvantage of this type of methodology is that it relies on non-testable modeling assumption. In reality, all of the previous solutions did as well to some degree. The advantage of structural estimation is that it can potentially answer many important questions, which cannot be answered using natural experiments.
12 Inequality

12.1 Characterizing the Distribution of Earnings

12.1.1 Measuring Inequality

Inequality is defined as the unequal distribution of resources in an economy. There are at least three broad ways to measure or define inequality in an economy.

1) Monetary Resources

We could examine inequality using a measure of monetary resources available to individuals and households, such as labor market earnings, income, and wealth.

2) Consumption

Another way to examine inequality would be to examine consumption inequality, such as the distribution of food, housing, or health care in an economy.

3) Outcomes

Still another way to examine inequality would be to look at the distribution of outcomes experienced by people, such as sickness, death, and happiness.
In this section we will focus primarily on the distribution of labor market earnings. Although there is generally a strong correlation between inequality in labor market earnings and other measures of inequality, this correlation is not perfect. It is important to note that different measures of inequality may provide distinct evidence about the nature of the distribution of resources within an economy.

12.1.2 Three Types of Monetary Resources

Labor Market Earnings

Earnings in the labor market context usually refer to wages, salary, bonuses, tips, and commissions earned by working. As we discussed earlier in the course, we should ideally include all forms of work compensation (including non-pecuniary benefits, pension benefits, health insurance, etc.) as part of labor market earnings. However, since these other forms of compensation are difficult to measure, we typically focus only on monetary earnings.

Income

Income is the flow of returns from either work or savings (investments). An individual’s total income in a period consists of all labor market earnings plus the returns to all other assets the individual owns (interest received from savings, stock dividends, etc.). For most working Americans, the majority of their income consists of labor market earnings.
Wealth

Wealth is the value of the stock of assets an individual owns at a point in time. Wealth at some period is the value of all the assets (houses, stocks, savings, etc.) the individuals owns minus the value of all debts. For many Americans, much of their wealth is embodied in their homes.

12.1.3 Three Concepts of Earnings Inequality

For the remainder of our discussion of inequality, we will measure inequality using labor market earnings (wages). Define labor market earnings for individual $i$ as $W_i$. I won’t specify a particular time dimension for these earnings (earnings could be measured hourly, weekly, monthly, annually, etc.)

We can define three distinct concepts of earnings inequality.

1) Cross-Sectional Inequality

The level of cross-sectional inequality indicates how unequally labor market earnings are distributed in an economy at a particular point in time. We can characterize cross-sectional inequality by the earnings distribution:

$$pr(W_{it} = w)$$

$i$ indexes individuals and $t$ indexes time. This function indicates the proportion of individuals in the economy at time $t$ with labor market earnings of level $w$. 
Earnings Differentials

Often it is useful to summarize the level of cross-sectional inequality with a single number. We’ll avoid the more complicated Gini coefficient in favor of a simple measure of inequality using the differential or ratio of earnings percentiles.

The median is the 50th percentile of earnings. We can also define other earnings percentiles the same way. In a population of 100 individuals, the 1st percentile is the earnings of the lowest earning individual. The 100th percentile is the earnings of the highest earning individual. The $P$th percentile of earnings is the earnings of the individual in which there are $P$ percent people earning less and $100 - P$ percent people earning more than that individual.

The 90-10 differential (as a ratio) is defined as

$$D_{90-10} = \frac{W_{90}}{W_{10}},$$

where $W_P$ is the $P$th percentile of earnings. $D_{90-10} = 1$ indicates no inequality as the earnings at the 90th and 10th percentiles are the same. $D_{90-10} > 1$ indicates inequality as earnings at the 90th percentile are greater than earnings at the 10th percentile.

The 50-10 differential or any other earnings differential is defined similarly. Wider differentials indicate a greater level of inequality in the economy.

In the United States today, $D_{90-10}$ is about 4.6. This indicates that
earnings at the 90th percentile are about 4.6 times higher than the earnings at the 10th percentile.

The 50-10 differential is much smaller. $D_{50-10}$ is about 2.2. The 50th percentile earns about 2.2 times more than the 10th percentile. This suggests that there is greater inequality at the top of the earnings distribution than at the bottom.

**College Premium**

Another simple measure of cross-sectional inequality is the *college premium*. Like the 90-10 differential, we’ll measure the wage gap between the college educated and the non-college educated (high school graduates) as a ratio:

$$\hat{\delta} = \frac{W_C}{W_H}$$

$\hat{\delta} = 1$ indicates that college and high school graduates earn the same wage. $\hat{\delta} > 1$ indicates that college graduates earn more.

Today, the college premium is about $\hat{\delta} = 1.9$. This indicates that the average college graduates earns 1.9 times more than the average high school graduate.

2) *Lifetime Inequality*

Lifetime inequality indicates the degree of social mobility in the economy. How earnings change with age (the difference in the earnings of young and old
workers) is a measure of lifetime inequality. To examine lifetime inequality, we can look at earnings as a function of age (the age-earnings profile we discussed in the Human Capital section):

$$ W_i = f(age_i) $$

We can write this as a regression model using log wages:

$$ \ln W_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \varepsilon_i, $$

where $age_i$ is individual $i$’s age in years.

As in the return to schooling section, the log function of wages allows us to interpret the population parameters in terms of percentage changes. For example, $\beta_1 = 0.05$ implies that an individual’s wage increase by 5 percent every year.

I estimated this regression model using the 2003 March CPS data sample. The OLS estimates are (standard errors in parentheses):

$$ \hat{\beta}_0 = 0.62 (0.028) $$

$$ \hat{\beta}_1 = 0.092 (0.0012) $$
\[ \hat{\beta}_2 = -0.0009 \ (0.00002) \]

These parameter estimates indicate that wages are increasing in age (\( \hat{\beta}_1 > 0 \)). The estimate of \( \beta_2 \) indicates that the age-earnings profile is concave.

How this function varies in the population with initial income can provide some evidence on social mobility. Do low-earning young workers have slower or faster wage growth over their lifetime than high-earning young workers? Define two different age-earnings profiles for low-earning and high-earning workers:

\[
W_{i}^{\text{low}} = \beta_{0}^{\text{low}} + \beta_{1}^{\text{low}} \text{age}_i + \beta_{2}^{\text{low}} \text{age}_i^2 + \varepsilon_{i}^{\text{low}}
\]

\[
W_{i}^{\text{high}} = \beta_{0}^{\text{high}} + \beta_{1}^{\text{high}} \text{age}_i + \beta_{2}^{\text{high}} \text{age}_i^2 + \varepsilon_{i}^{\text{high}}
\]

Assume \( \beta_{0}^{\text{low}} < \beta_{0}^{\text{high}} \). If the two earnings growth rates are the same (\( \beta_{1}^{\text{low}} = \beta_{1}^{\text{high}} \)), then there will be no convergence in earnings over the lifetime. If \( \beta_{1}^{\text{low}} < \beta_{1}^{\text{high}} \), then the gap between rich and poor increases over the lifetime.

3) Intergenerational Inequality

Intergenerational inequality measures the persistence of inequality across generations. This is typically measured by estimating an intergenerational
earnings elasticity. Let $W_i^p$ be the earnings of individual $i$’s parents (in most studies, this is the father’s earnings). $W_i$ is the wage of the child (in most studies, this is the son). The regression model is

$$\ln W_i = \beta_0 + \beta_1 \ln W_i^p + \varepsilon_i$$

(Note: Because earnings change with age, we would want to compare the earnings of parents and children at the same age, e.g. earnings when the father and son are both age 35.)

If $\beta_1 = 0$, there is no relationship between parents’ and children’s earnings. A positive and large $\beta_1$ indicates that there are substantial transfers between generations. These transfers could include wealth (e.g. wealth transfers lower credit constraints for the children’s human capital investments) or ability (e.g. smart parents have smart kids).

Several studies have estimated $\beta_1$ to be $\hat{\beta}_1 = 0.4$ or higher for the United States.

$\beta_1$ indicates the extent to which inequality is transferred across generations. If the parents have 50 percent higher labor market earnings than the population average of the parents’ generation, with $\hat{\beta}_1 = 0.4$, the child is expected to have earnings $50 \times 0.4 = 20$ percent higher than the average in her generation. The grandchildren are expected to have $50 \times 0.4 \times 0.4 = 8$ percent higher earnings than the average in the grandchildren’s generation.

It is important to note that intergenerational inequality could influence
the level of cross-sectional inequality through credit constraints to finance
human capital. The level of intergenerational transfers may determine the
extent to which individuals are credit constrained in their human capital
investments.

12.2 What Determines the Level Cross-Sectional In-
equality in the Economy?

For the remainder of this section, we will focus on cross-sectional inequality.
First, let’s briefly consider three major factors that determine the level of
cross-sectional inequality in an economy.

1) Differences in Human Capital Levels

As we have discussed previously, a major determinant of wage levels is
human capital. Let’s return to our simple model of wages and schooling.
Assume there are two groups: college educated and non-college educated.
Their wages in period $t$ are $W_{tC}$ and $W_{tH}$. Assume $W_{tC} > W_{tH}$.

The distribution of earnings in this simple economy is determined by the
proportion of the population with a college degree ($p_{Ct}$). We can write the
cross-sectional distribution of earnings in period $t$ as

$$
pr(W_{it} = W_{tC}) = p_{Ct},
$$
\[ \text{pr}(W_{t \propto} = W_{t \propto N}) = 1 - p_{Ct}, \]

We can close the gap in earnings \((W_{t \propto} - W_{t \propto H})\) in two ways:

i) **Convergence in Human Capital Investments**

If all individuals become college educated \((p_{Ct} = 1)\), then all individuals earn \(W_{Ct}\).

ii) **Convergence in Human Capital Returns**

If the college premium is eliminated \((W_{t \propto C} = W_{t \propto H})\), then all individuals earn the same wage.

2) **Government Taxes and Transfers**

Another important determinant of cross-sectional inequality is the extent of government taxes and transfers. In principle (although often not in practice), the government can act to redistribute income and wealth from the rich to the poor.

For a two group model, the government can equalize labor market earnings by taxing the college educated \(T\) and transferring this income to the non-college educated. The after tax wage that individuals take home is now \(W_{t \propto C} - T\) for the college educated and \(W_{t \propto H} + T\) for the non-college educated.
As with most tax systems, the tax imposes a distortion in the economy. In the extreme case in which the tax $T$ is so large as to force $W_{tC} - T = W_{tH} + T$, the monetary return to college is zero, and it would be unlikely that many individuals would incur the cost of schooling and obtain a college degree.

3) Discrimination

Another potential determinant of the level of cross-sectional inequality is discrimination. Discrimination in the labor market is discussed in more detail below. Briefly, discrimination can act as a “tax” on the earnings of some groups and widen the gap in earnings directly. Discrimination can also lower the return to human capital investments for some groups and thereby widen the difference in human capital levels in the economy.

12.3 Trends in Cross-Sectional Earnings Inequality

12.3.1 Trends

Trends in cross-sectional inequality have received a lot of attention in economics. The general pattern since World War II was a decline in inequality up until the 1970s. In particular, the period of the 1940s and 50s has been termed the “Great Compression.” Since the late 1970s, inequality has increased substantially.

Let’s look at trends in inequality measured using the college premium ($\hat{\delta}$) defined above.
In 1970, $\hat{\delta} = 1.6$. In 1980, $\hat{\delta} = 1.5$. In 1990, $\hat{\delta} = 1.7$. In 2000, $\hat{\delta} = 1.9$. Over the 20 years from 1980 to 2000, the college premium nearly doubled. This rapid rise in inequality has attracted a considerable amount of research.

Similar trends in the 90-10 differential occurred. In 1970, $D_{90-10}$ was about 3. By 2000, $D_{90-10}$ was 4.6.

12.3.2 Explaining the Trends

Relative Supply and Demand

One potentially useful way to explain these trends is to examine the market for skill. We can think of the college and non-college workers as supplying two distinct types of labor or two types of labor skill. The college and non-college groups comprise two skill groups, where the college group is considered “skilled” and the non-college group is “unskilled”. The college premium can then be thought of as a skill premium.

Let’s construct a supply and demand graph for relative skill. On the vertical axis, we measure the relative skill price (relative wages) or skill premium: $\frac{W_C}{W_H}$. On the horizontal axis, we measure the relative labor employed from the two groups: $\frac{H_C}{H_H}$.

The curves in this graph represent relative labor supply and demand. The relative labor supply curve indicates how the supply of college educated labor relative to high school educated labor responds to changes in the relative wages. The relative labor demand curve indicates how the demand from firms
for college educated labor relative to high school educated labor responds to changes in the relative wage. As with any labor supply and demand curves, relative labor supply slopes upward and relative labor demand slopes downward.

*Shifts in the Relative Supply of Skill*

The stock or supply of college graduates and non-college graduates in any particular year is composed of all the individuals who are now of working age (18-65 say). The current stock of college graduates and high school graduates depends on three factors:

i) the size of previous birth cohorts (flow of new workers),

ii) the fraction of individuals from each birth cohort who graduated from college (flow of new college educated workers), and

iii) the number of college and non-college educated immigrants.

One reason for the low level of the skill premium in the 1970s is that during this period there was an increase in the number of college educated workers. This was due to the high numbers of individuals from the “baby boom” birth cohort who graduated from college and entered the labor market in the 1970s.

A supply-side explanation of the increase in the college premium from 1980 to 2000 requires a shift in of the relative supply curve over this period. In
fact, the opposite occurred. The percent of college graduates in the workforce increased from 21 percent in 1980 to 28 percent by 2000. If the relative labor demand curve remained fixed over this period, this shift out of the relative supply curve should have reduced the college premium.

Immigration may have also played a role in increasing inequality in the United States, especially for the lower half of the earnings distribution (which is measured by the 50-10 differential). The number of immigrants entering the United States over the 1966-2000 period (either legal or illegal) was sizable (over 30 million). These immigrants had low levels of education on average and they increased the number of high school dropouts much more than the number of college graduates.

*Shifts in the Relative Demand for Skill*

Given this shift out in the relative supply curve, economists have turned to shifts in of the relative demand curve as an explanation for the 1980-2000 increase in the skill premium. The relative demand curve for skill represents the relative aggregate demand for skill by all firms in the United States. Two basic factors can cause the the relative supply curve to shift out:

i) a relative reduction in the output price for goods produced by firms that employ relatively more unskilled labor, or

ii) an increase in the relative productivity of skilled labor versus unskilled labor (*skill biased technological change*).
Most firms hire both skilled and unskilled labor. However, firms in certain industries hire more of one type than the other. Manufacturing firms (e.g. firms manufacturing steel, automobiles, etc.) generally hire more unskilled workers than many service firms (e.g. firms producing health care, education, or financial services). In addition, unskilled workers in manufacturing firms have higher wages on average than unskilled workers in firms in other industries.

It has been argued that one of the reasons the skill premium increased over the 1980-2000 period is that international competition decreased the output price of manufactured goods relative to the output price of other goods (e.g. service goods). As we studied in the labor demand section, a reduction in the output price causes firms to reduce labor demand. The reduction in the price of manufactured goods in the United States caused the relative demand curve for skill to shift out.

Economists have also argued that that skill biased technological change shifted the relative demand curve for skill out. Over the 1980-2000 period, computers, manufacturing automation, and other technologies increased the productivity of skilled workers relative to unskilled workers. As the relative productivity of skilled workers increased, the relative demand curve for skilled workers shifted out.
12.4 Earnings Differences for Sub-Groups

12.4.1 Gender

The well known *gender gap* in earnings can be measured by the ratio of average earnings for women versus men:

\[ \hat{G}_{wm} = \frac{W_{\text{women}}}{W_{\text{men}}}. \]

A value of \( \hat{G}_{wm} = 1 \) indicates men and women have the same average earnings. A value of \( \hat{G}_{wm} < 1 \) indicates women are paid on average less than men.

In 2000, the gender gap was about \( \hat{G}_{wm} = 0.75 \). Average women’s wages were 75 percent of average men’s wages. This gender gap has declined over time, especially after 1980. In the 1960s, the gender gap was larger at \( \hat{G}_{wm} = 0.6 \).

12.4.2 Race

Similar to the gender gap, we can define several racial gaps in earnings. Let’s measure the gap in earnings between blacks and whites using this ratio of average earnings:

\[ \hat{G}_{bw} = \frac{W_{\text{black}}}{W_{\text{white}}}. \]

A value of \( \hat{G}_{bw} = 1 \) indicates blacks and whites have the same average
earnings. A value of \( \hat{G}_{bw} < 1 \) indicates blacks are paid on average less than whites.

In 2000, the black-white earnings gap was about \( \hat{G}_{bw} = 0.8 \). Average black wages are 80 percent of average white wages. This black-white earnings gap has declined somewhat over time. In the 1960s, the black-white earnings gap was larger at \( \hat{G}_{wm} = 0.7 \).

### 12.5 Explaining Earnings Differences between Sub-Groups

The two competing explanations for the gender and black-white earnings gaps are i) human capital differences and ii) discrimination. To the extent that these groups have different levels of human capital, these earnings gaps may be due to these human capital differences. The discrimination explanation posits that even if individuals have the same level of human capital and productivity, the wages of women and minorities are lower because of discrimination. Another way to express this is that the return on the human capital for women or minorities is lower than the return on the human capital for men and whites.

These two explanations may not be unrelated. As I discuss in more detail below, discrimination in the labor market may cause groups discriminated against to reduce their investment in human capital.
12.5.1 Decomposition Analysis

One potentially useful approach to examining differences in sub-group earnings is a decomposition analysis. In this analysis, we want to decompose the observed earnings differences into a part that can be “explained” by differences in measurable human capital levels and a part that cannot be explained by human capital.

For simplicity, again assume there are two levels of human capital: college and non-college. The expected wage for any two groups (A and B) is given by

\[ E[W_A] = E[W_A|\text{college}] \cdot P^A_{\text{col}} + E[W_A|\text{high}] \cdot (1 - P^A_{\text{col}}), \]

\[ E[W_B] = E[W_B|\text{college}] \cdot P^B_{\text{col}} + E[W_B|\text{high}] \cdot (1 - P^B_{\text{col}}), \]

where \( E[W_j|\text{college}] \) is the expected wage of college graduates for group \( j = \{A, B\} \), \( E[W_j|\text{high}] \) is the expected wage of non-college graduates, \( P^j_{\text{col}} \) is the proportion of the population with a college degree, and \( (1 - P^j_{\text{col}}) \) is the proportion of the population without a college degree.

We can write the difference in the expected wages between the two groups as

\[ E[W_A] - E[W_B] = E[W_A|\text{college}] \cdot (P^A_{\text{col}} - P^B_{\text{col}}) + \eta, \]
where $\eta$ is are all the other remaining terms.

The two parts of the difference in expected wages is

i) the difference in human capital levels ($P_{col}^A - P_{col}^B$), evaluated at the expected college graduate wage for group $A$, $E[W_{A|college}]$, and

ii) the residual difference represented by $\eta$.

*Estimating the Decomposition*

With a random sample, we can estimate each of the components in the decomposition.

Let’s examine the difference in average wages between black and white workers. Here are our descriptive statistics (wages are hourly wages):

$$\bar{W}_{black} = 16.80$$

$$\bar{W}_{white} = 21.00$$

$$\hat{P}_{col}^{black} = 0.17$$

$$\hat{P}_{col}^{white} = 0.27$$

$$\bar{W}_{white,col} = 28$$

The decomposition is

$$21 - 16.8 = 28 \times (0.27 - 0.17) + \eta$$
The decomposition indicates that about $\frac{2}{3} \cdot (2.8/4.2 = 2/3)$ of the difference in black and white wages can be attributed to differences in human capital levels. The remaining $\frac{1}{3}$ is attributed to $\eta$, which arguably reflects discrimination.

### 12.5.2 Problems with Decomposition Analysis

Should we believe that only $\frac{1}{3}$ of the difference in black and white wages is due to discrimination ($\eta$)? The decomposition analysis is intended to answer the following question: What would the black-white wage gap be if blacks had the same level of human capital as whites?

However, as we have discussed at several points, human capital levels (e.g., college degrees) are not randomly assigned. If discrimination in the labor market reduces the return to a college degree for blacks relative to whites, then part of the difference in human capital levels should also be attributed to discrimination. That is, our estimate that 87.5 percent of the difference in black-white wages is due to differences in human capital is an over-estimate.

Another important caveat to the decomposition analysis is that it only uses observed human capital. There may be many other differences in un-
observed human capital between blacks and whites and men and women. In several studies, economists have found that including more measures of other forms of human capital (e.g. tenure on the job, occupation, college major, etc.) reduces differences in earnings even further. However, the same problem arises here as with schooling human capital. Discrimination in the labor market may be reducing the returns to these other forms of human capital as well and thereby causing some of the differences in observed human capital levels.

12.6 Discrimination

12.6.1 Taste Discrimination

Taste discrimination is the result of a preference for one group over another. Taste discrimination occurs when someone receives a higher utility (non-pecuniary benefits) from interactions with members of a particular group (e.g. whites prefer to interact with whites). This prejudice can take several forms:

i) Employers prefer employing some workers over others.

ii) Workers prefer working with some co-workers over others.

iii) Customers prefer purchasing products from some workers or some firms over others.

Taste Discrimination as a “Tax” on Wages
The key aspect of taste discrimination is that the prejudice stems from a non-pecuniary preference. But this non-pecuniary preference is reflected in wage rates. Take the example of taste discrimination of whites against blacks. The preference of employers, co-workers, or customers for white workers essentially imposes a “tax” on the wage of black workers relative to white workers. This tax lowers the demand for black workers relative to white workers. The tax shifts the demand curve in for black workers and pushes out the demand curve for white workers. Fewer black workers are employed and their wage is now lower than that of white workers.

A Simple Model of Taste Discrimination

Employers, co-workers, and customers receive utility from workers. This utility is composed of a pecuniary part (wages) and a non-pecuniary part (their preferences for them). Utility from black workers and white workers is

\[ U_b = W_b - D, \]

\[ U_w = W_w \]

where \( D > 0 \) reflects the taste discrimination. \( D \) is the tax on black worker earnings. For black workers to be hired, the utility others receive from them must be equal to the white utility: \( U_b = U_w \). This implies that the black wage must be lower than the white wage by \( D \):
$W_b = W_w - D$

Compensating Differentials

The effect of taste discrimination on wage differentials can be viewed within the context of the compensating differentials model we studied earlier. Black workers must compensate others for their lower level of non-pecuniary benefits they provide through offering lower wages. They essentially have to “bribe” employers through lower wages in order to get hired.

12.6.2 Statistical Discrimination

Statistical discrimination is where employers use group stereotypes to determine wage offers. We already discussed a type of statistical discrimination in our discussion of signalling theories of human capital. In that model, employers had imperfect information about worker productivity. Firms used an observable characteristic, observed schooling, to infer whether an individuals was a high productivity (high ability) type or a low productivity (low ability) type. Wages were determined by the level of schooling the individual possessed.

In the same way, employers could use any number of observable characteristics to determine wage levels: gender, race, height, weight, age, etc. The
major difference between the signalling and statistical discrimination models is that unlike schooling in our signalling model, these physical characteristics are immutable.

A Simple Model of Statistical Discrimination

Assume there are two groups: group $A$ and group $B$. All firms know the average productivity of both groups, $q_A$ and $q_B$. Assume group $A$ is more productive: $q_A > q_B$. If firms know nothing else about the workers, they will offer the following wages. A worker belonging to group $A$ receives $W_A = q_A$ and a worker belonging to group $B$ receives $W_B = q_B$.

In this model, the high productivity workers in group $B$ are discriminated against because they belong to the low average productivity group $B$. In general, if a worker is below average relative to her group, she benefits. If a worker is above average relative to her group, she loses.

This model explains gender and racial earnings differences as a lack of information by firms on the true productivity of individual workers. Women and minorities receive lower average wage offers by firms because on average firms believe they are less productive.

We should not interpret the lower perceived average productivity of women and minorities as reflecting genetics. Instead, it may reflect lower average levels of unobserved human capital for these groups relative to men and whites. In the case of women, lower perceive average productivity may reflect the perception that women will be less likely to stay with a firm and make in-
vestments in firm training (e.g. women are more likely to take time off to take care of their children or a sick parent).

12.7 Affirmative Action

Since the civil rights era in the 1950s and 1960s, most forms of institutionalized discrimination have been declared illegal in the United States, and substantial resources have been devoted to the enforcement of these laws. Gender and racial earnings gaps have not fully closed, however. Since the 1960s, additional policies, called affirmative action policies, have been implemented in an attempt to reduce the remaining inequities. These laws take several forms, including racial preferences for college admissions and requirements that a certain percentage of government contracts be filled by firms who are owned by women or minorities or firms that employ a sufficient number of women and minorities.

Affirmative action policies are very controversial. Let’s briefly discuss some major advantages and disadvantages of these policies.

12.7.1 Advantages

1) Lower Earnings and Employment Differences Directly

Affirmative action programs directly transfer resources to women and minorities through education subsidies and improved labor market conditions. This can result in a direct reduction in earnings inequality.
2) *Change Perceptions of Employers, Co-Workers, and Customers*

Increased employment for women and minorities in firms, occupations, and industries where their representation was previously small may help to dispel inaccurate stereotypes. This could change the perceptions of employers, co-workers, and customers and reduce discrimination.

3) *Increase Returns to Human Capital*

Greater labor market opportunities could increase the return to human capital investments and cause women and minorities to increase their levels of human capital investments. By increasing the number of women and minorities in better paying jobs, affirmative action could also help create important role models and job networks for women and minorities.

4) *Diversity as an Externality*

To the extent that diversity benefits others (e.g. affirmative action in college admissions improves the learning environment), affirmative action programs may also produce positive externalities.

### 12.7.2 Disadvantages

1) *Mis-Allocate Resources*

If we assume that prior to affirmative action programs the labor market is operating efficiently, then any intervention in the labor market through
affirmative action policies may be a mis-allocation of resources and be inefficient. For example, it has been argued that affirmative action programs in schooling shift educational resources to less qualified minority applicants.

2) *Undermine Positive Gains*

Affirmative action programs may create the perception that the accomplishments of women and minorities are undeserved. Affirmative action may create new stereotypes and increase discrimination, and thereby undermine the gains under-represented groups have made.

3) *Ineffective without Prior Human Capital Investments*

As we discussed in the human capital section, complementarities in human capital production imply that later interventions are less effective in changing outcomes than early interventions. Affirmative action programs in the labor market and post-secondary education may be less effective given that women and minorities have lower levels of prior human capital. A more effective public policies is to use subsidies to equalize early human capital differences.