Excerpts from

Chicago Price Theory

Comments are welcome: please email to

supplyanddemand@uchicago.edu

Part II: Market Equilibrium

Chapter 7 Discrete Choice and Product Quality

Imagine a setting in which individuals are deciding whether to buy either 0 or 1 of a good. Consider Figure 7-1.

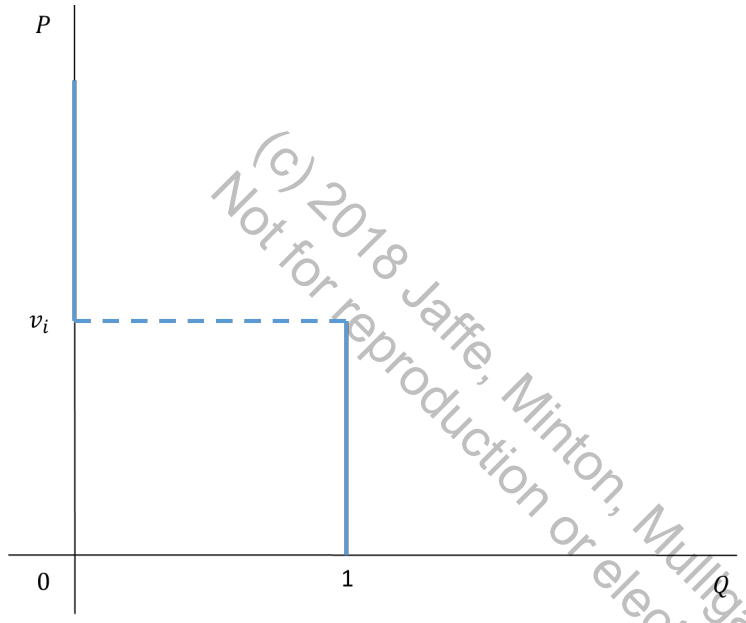


Figure 7-1: At v_i, one is indifferent between buying the good or not.

Here we are ignoring that a lot of goods that look discrete can be thought about as more continuous. Multiple haircuts are purchased over time, for example.

Market Demand is a Distribution Function

Each person has a cutoff value v_i, which gives the value they place on the good in dollars. At v_i, individual i is indifferent about buying the good. This can be thought about empirically as well. We can infer v_i, for instance, by looking for the price at which a person moves from not buying to buying.

We can further consider a distribution $F(v) = \Pr(v_i \leq v)$, which gives the fraction of the population with a value less than v. Then the demand at a price P will be given by $D(P) = (1 - F(P))N$, where N is the number of people in the population and $1 - F(P)$ is the fraction of people who buy.

Suppose we consider a normal distribution. What would the demand curve look like? See Figure 7-2. The function asymptotes toward N, which is the maximum number of units that could be sold.

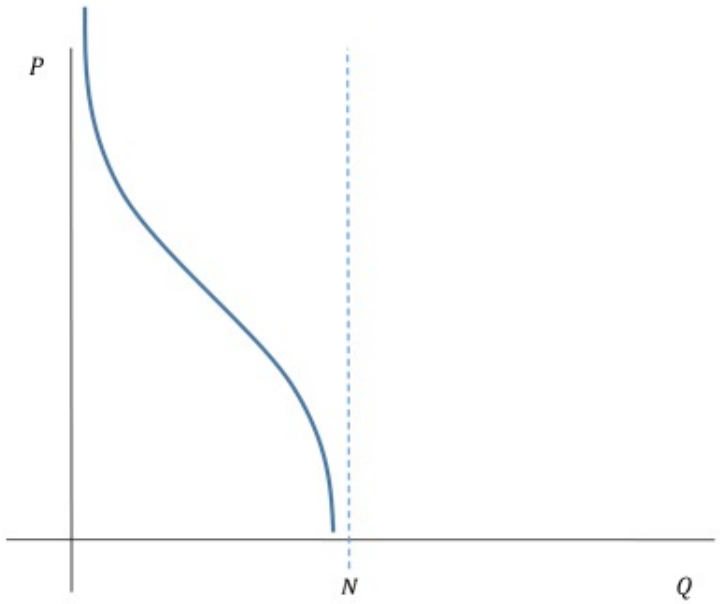


Figure 7-2: Normal distribution case.

What if the distribution is uniform? That is, $V \sim U(A, B)$. At B, no one buys, and at A, everyone buys. The demand curve is depicted in Figure 7-3. Note that it is linear. Linear demand curves are often motivated as representing a uniform distribution of preferences.

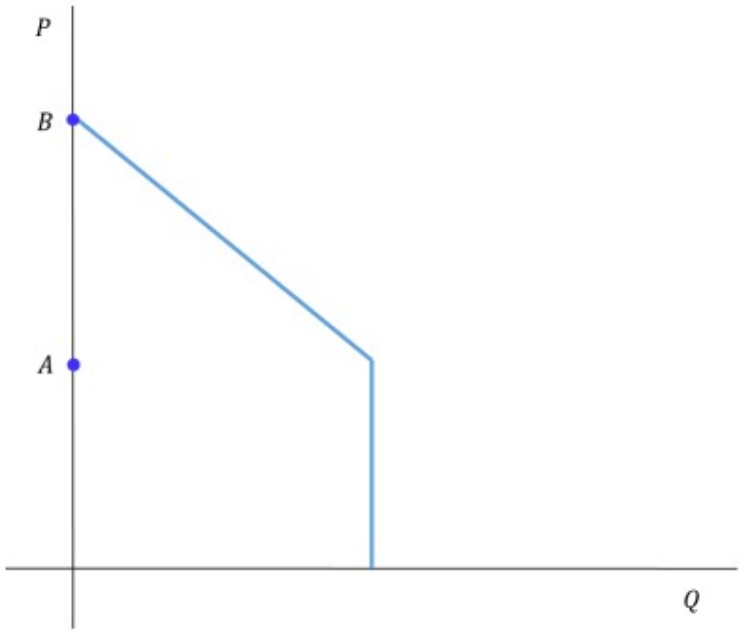


Figure 7-3: Uniform distribution case.

The elasticity of demand for a general distribution is given by

$$\frac{P}{D(P)}\frac{\partial D(P)}{\partial P} = -\frac{PNf(P)}{\big(1-F(P)\big)N} = -\frac{Pf(P)}{1-F(P)}$$

where $f(P) = F'(P)$ is the probability density function corresponding to F.

The elasticity of demand depends on the fraction of people on the margin versus the fraction of people already buying. At a point with high density, demand will be relatively elastic. We can think about these demand functions as adoption curves; that is, as new goods come out, these curves tell us how many people will buy as the price decreases over time. Suppose appliances are getting cheaper; consider the demand curve for the normal distribution in Figure 7-2. Once the price gets low enough, demand is very elastic, so additional reductions in price cause large changes in purchasing. The fact that the middle class goes on a spending spree when appliances get cheap enough isn't necessarily a bandwagon or network effect. For cell phones, this model might not be enough. Cell phones might additional require a network effect explanation because more people owning cell phones raises non-owners' desires to own cell phones.

Now note that whenever we have a demand curve, we can approximate it using a linear demand curve. This type of procedure is not always sufficient for the analysis we might want to do—in monopoly problems, for instance, often the curvature of the demand curve can be very important. For many purposes, however, if we're interested in the price-quantity relationship, honing in on a small area will be useful. See Figure 7-4. The question is about whether we're approximating all the elements relevant for behavior. For assessing welfare, for instance, we can look at the budget constraint as an approximation to an indifference curve. If we're analyzing substitution, however, a linear approximation of the indifference curve means we get perfect substitutes. So for this, we'd have to model the curvature of the indifference curve.

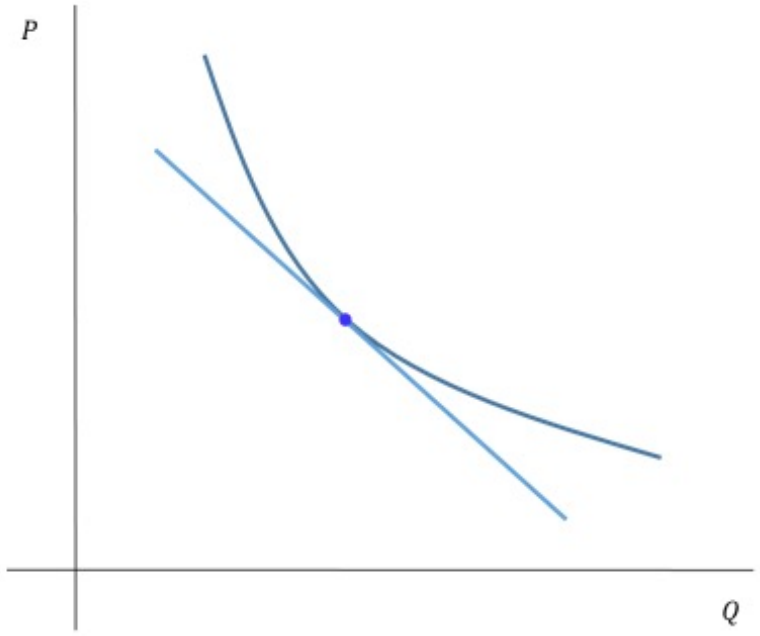


Figure 7-4: Approximating the demand curve with a linear demand curve.

Equilibrium Product Quality

Once we're operating in the world of single-choice, we can solve more complicated problems. Now we will stay focused on the discrete choice but extend the model to think about varying levels of quality. Let

q = quality. People will still just buy 1 unit of the good, such as a TV, but now they will also consider what degree of quality they want in their TV. For simplicity, we will assume quality is continuous. For the TVs, if we think about size, this means you can buy a 59.1 inch TV. This isn't particularly realistic but will make our lives easier.

We'll consider indifference curves $U(x, q)$, where x denotes other goods, $P_x = 1$, and $M =$ Income. So $x = M - P$, where P is the price of the TV. This means we can rewrite the utility function

$$\overline{U}(M - P, q)$$

Sometimes we assume quasilinear utility $U(x, q) = x + V(q) = M - P + V(q)$. This means there is a constant marginal utility of income, since utility is linear in x.

Pick q to max $V(q) - P(q)$, where $P(q)$ is the schedule showing the price for each quality level. This means quality is not a normal good under quasilinear utility—M doesn't matter for the choice of q. Though quasilinear is a popular model, we know quality choice increases with income. We can use a more general utility function $\bar{U}(M - P, q)$ to get quality to be a normal good.

Why do people increase quality as income goes up (rather than quantity)? There are physical constraints—stomach capacity, time in the day, etc. How might we model this quantity-quality problem? Consider

$$U\big(X, NV(q)\big) + \lambda[M - X - NP(q)]$$

Where N denotes quantity. We can rewrite this as

$$U(X, Z) + \lambda\left[M - X - Z\frac{P(q)}{V(q)}\right]$$

Where $Z = NV(q)$ denotes "effective consumption," since it summarizes both quantity and quality. There's a q^* that marks the efficient quality level (the ratio $\frac{P(q)}{V(q)}$ denotes cost per unit enjoyment). The world puts some limits on these parameters. For example, we may have $N \leq \overline{N}$. $\overline{N}$ could be the capacity of your stomach, for instance. Then as income rises, people will increase quality. This might explain why poor people don't just buy less of the high quality goods.

Now consider a slightly more general model, where q_j = quality of j and N_j = consumption of j. We can write the utility function

$$U\left(\sum_{j=1}^{J} N_j, \sum_{j=1}^{J} N_j q_j, X\right)$$

In this utility function, an individual cares about total amount and quality weighting. An interesting issue is whether $\sum_{j=1}^{J} N_j$ has positive or negative utility. Implicitly, it has negative marginal utility, since it's a constraint (limited by $\overline{N}$). Maybe the goods are various types of foods, and quantity is measured by they calories they deliver and quality is measured by their taste or eating experience. People enjoy eating but do not want too many calories, at least if their income is high enough that starvation is not a concern. In general, so long as demand for $\sum_{j=1}^{J} N_j$ grows slower than $\sum_{j=1}^{J} N_j q_j$, you'll move up the quality ladder as you consume more.

In general, people prefer lower prices and higher quality. See Figure 7-5. There is no reason the curves here need to be concave, but let us suppose they are. How will we describe consumer behavior in this model?

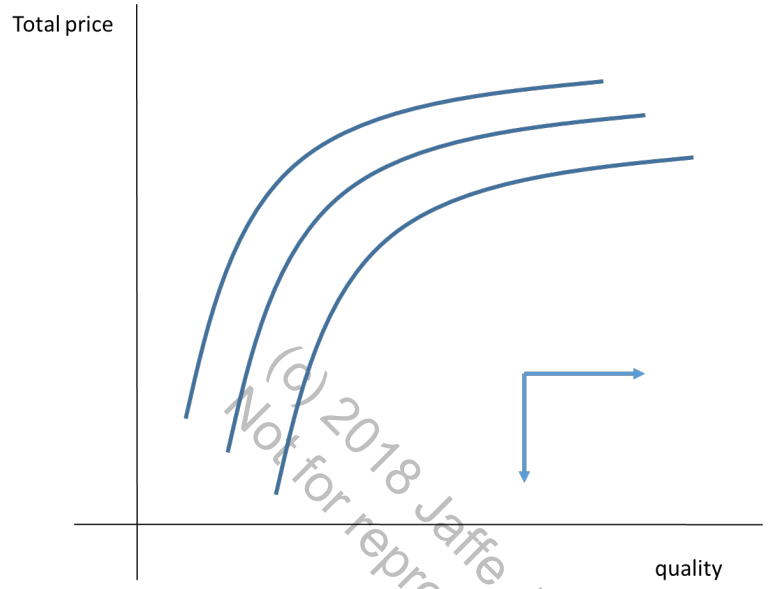


Figure 7-5: Consumers prefer higher quality and lower price.

When the consumer goes to the store to buy his TV set, we will consider a price $P(q)$, a price increasing in quality q. The consumer will choose a point where his indifference curve is tangent to this price line from below. See Figure 7-6.

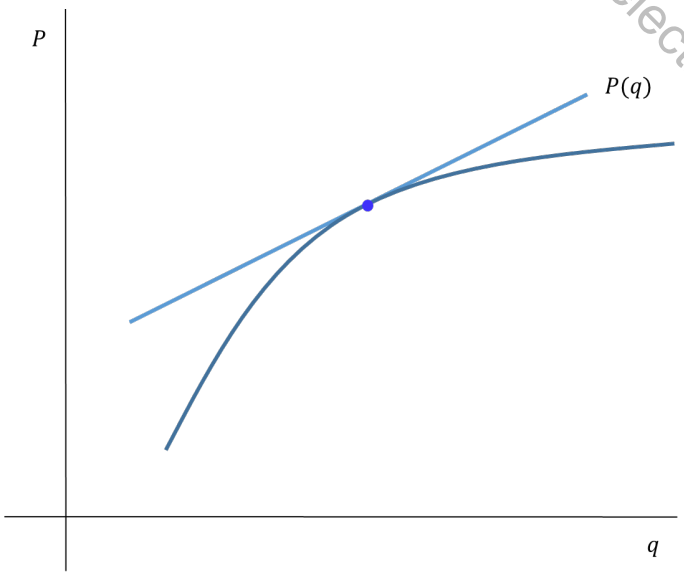


Figure 7-6: Slope gives us the marginal willingness to pay for an additional unit of quality.

The indifference curve always has to be concave relative to the equilibrium price line, which is why it did not matter that we assumed indifference curves were concave. (If part of the indifference curve is more convex than the price line, the consumer will never choose that point in equilibrium.) The slope near the tangency will give us the marginal willingness to pay for an additional unit of quality.

Now let's think about the firm side of the market. Assume there are a large number of producers and that the unit cost of production is $C(q)$. Each producer makes one unit and chooses the quality to produce. Let $N =$ number of consumers and $M =$ number of producers. Assume $M > N$. This implies profits $\Pi = 0$. Why? Some producers are not going to produce in equilibrium, which means they earn 0 profit. This means the producers who are producing must be making 0 profits. Thus $P(q) = C(q)$.

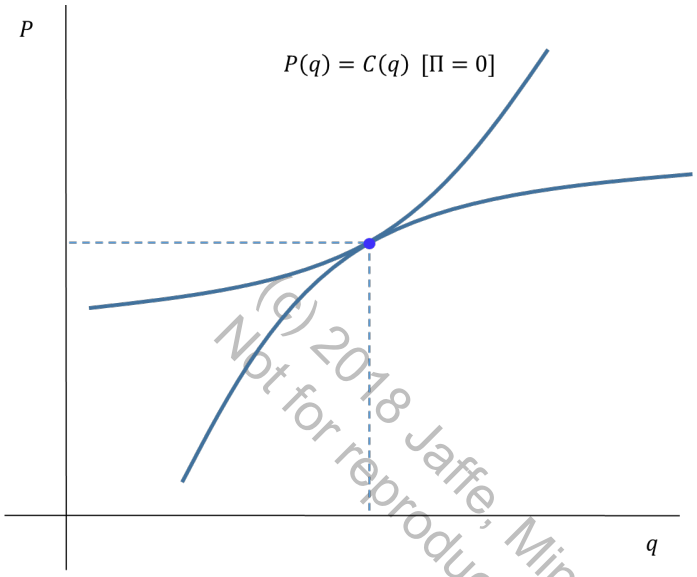


Figure 7-7: Consumers pick a level of quality where their indifference curves are tangent to the price curve.

Now, as before, consumers pick a level of quality where their indifference curves are tangent to the price curve. See Figure 7-7.

What happens if someone gets richer? Their indifference curves will get steeper, since they will have a higher preference for quality, and quality is a normal good. They will cross the indifference curves of the poorer consumers from below. See Figure 7-8.

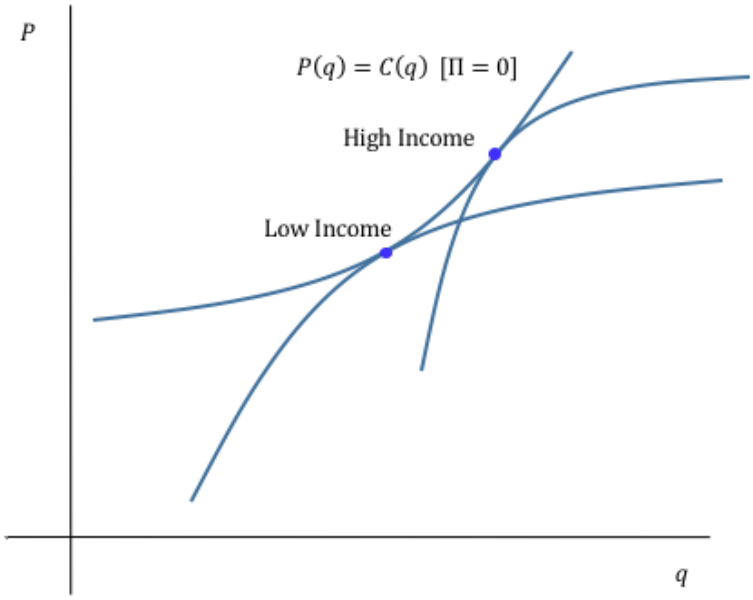


Figure 7-8: Consumers pick higher levels of quality as they become richer, since q is normal.

Heterogeneity among consumers pushes them to consume along different points along the producer's cost curve. The price differentials at each point represent each consumer's marginal willingness to pay for quality and also trace out the producer's overall marginal cost of quality.

Let's think a little more about what q being normal implies about indifference curves. Consider Figure 7-9, and pick a point on the indifference curve. Then as we increase vertically from this point, it must be that the indifference curves are becoming steeper, because this is what would drive additional consumption of quality. That is, q normal implies that willingness to pay for q rises as X increases, holding q fixed. The normality condition for X would be the reverse. As one moves right along a horizontal line from the chosen point along the indifference curve, the indifference curves must be getting flatter, since this would drive increased consumption of X. Note that this is true in the two good case but more complicated in the multi-good case.

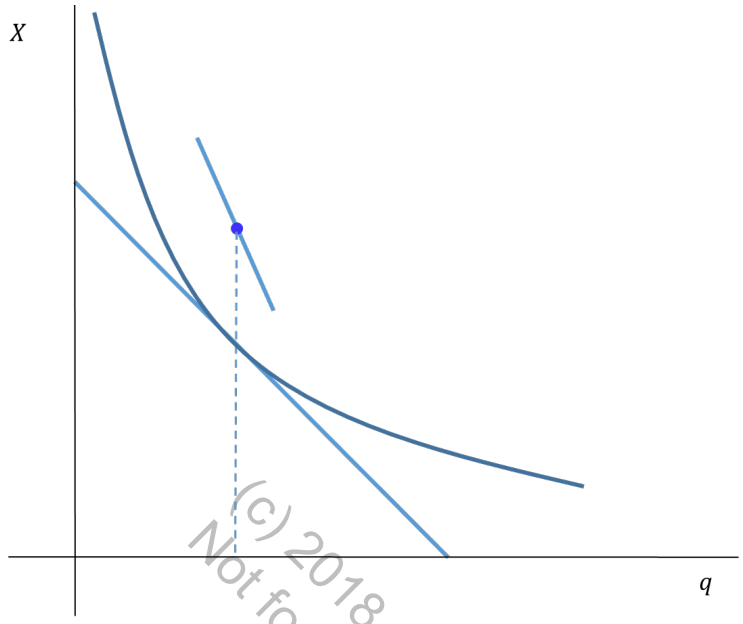


Figure 7-9: If q is normal, then the willingness to pay for q rises as X increases, holding q fixed.

Thus, normality of q means that

$$\frac{\partial(U_q/U_X)}{\partial X} > 0$$

Heterogeneous Firms

Consider again the quality indifference curves for two consumers, depicted in Figure 7-10.

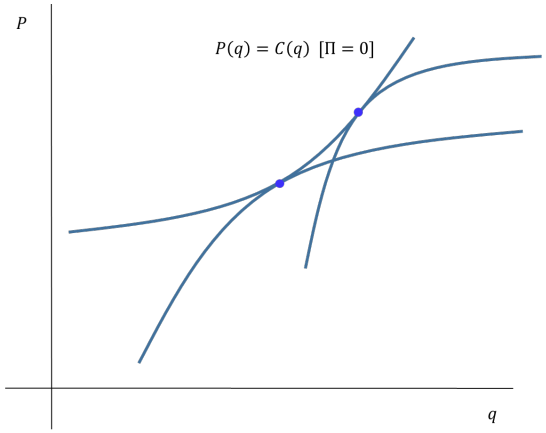


Figure 7-10: Different allocations along the quality cost curve.

Different consumers sorted out along the cost curve. This was the case of two different consumer types and one firm type. Now what if we change the problem to consider two different types of firms and one type of consumer? Assume free entry of both types of firms, so there are many firms of type 1 and many firms of type 2. Would we see more than one point in equilibrium? See Figure 7-11. Unless it is a knife-edge situation, where the price of additional quality consistent with zero profits exactly coincides with consumer willingness to pay for quality, we would end up with a single point. In other words, one of the types of firms is offering a better deal than the other, which cannot provide consumers with the same

utility with taking a loss. Profits of the former firms would be driven to 0, and consumers would buy from the firm that better satisfied their needs. Free entry of firms guarantees this.

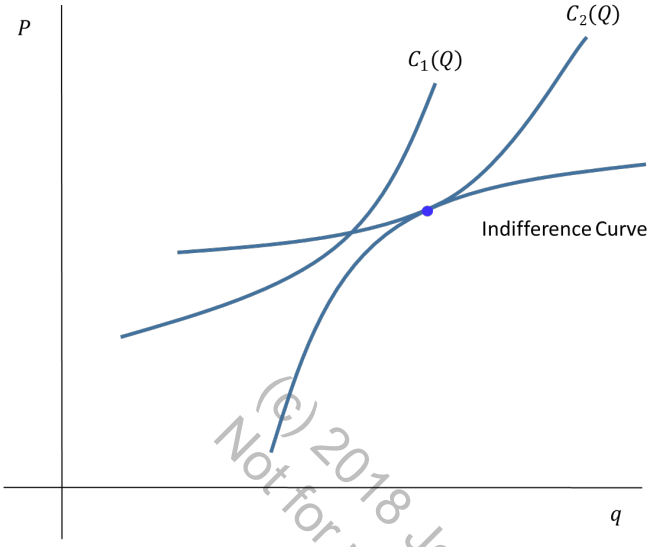


Figure 7-11: Ignoring capacity constraints (for the moment), homogeneous consumers buy from firms of type 2.

What happens if there is a limited supply of firms? Suppose $N_1 + N_2 > N_{cons}$, $N_1 < N_{cons}$, and $N_2 < N_{cons}$. So there are less firms of each type than total consumers, but both firms together can produce to satisfy all the consumers (we're still in the discrete production case). There will be a unique equilibrium. People prefer to consume from firm 2, but firm 2 cannot satisfy everyone. So all firms of type 2 will produce, but some type 1 firms will be required. But since $N_1 + N_2 > N_{cons}$, type 1 firms must earn 0 profits. No firm of type 1 can make profits in equilibrium, otherwise other firms would enter. Firms of type 2 are constrained to offer a combination that's at least as good as the offerings by the type 1 firms. So they produce along the curve where they are also tangent to the consumer's indifference curves

In other words, we draw the equilibrium by first tracing out the combinations of price and quality that yield zero profit for type 1, which is the cost curve $C_1(q)$. See Figure 7-12. Then we draw the indifference curve tangent to it, because any type 1 firm deviating from the tangency point must either take a loss or get no customers (who prefer to take the tangency point offered by other type 1 firms).

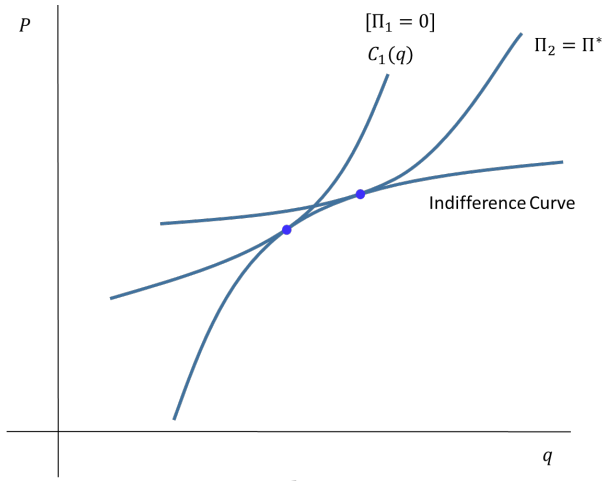


Figure 7-12: Firms of type 1 make no profit in equilibrium. Firms of type 2 produce according to $\Pi_2 = \Pi^*$ to maximize profits.

Finally, the type 2 firms maximize profits subject to the constraint that consumers are no better off buying from a type 1 firm instead. Note that the $\Pi_2 = \Pi^*$ curve is not its cost curve, but is above it because they are able to charge more than cost.

Consider more generally what the firm's indifference curves look like, depicted in Figure 7-13. The positive profit curve in Figure 7-12 is just the cost curve shifted upwards.

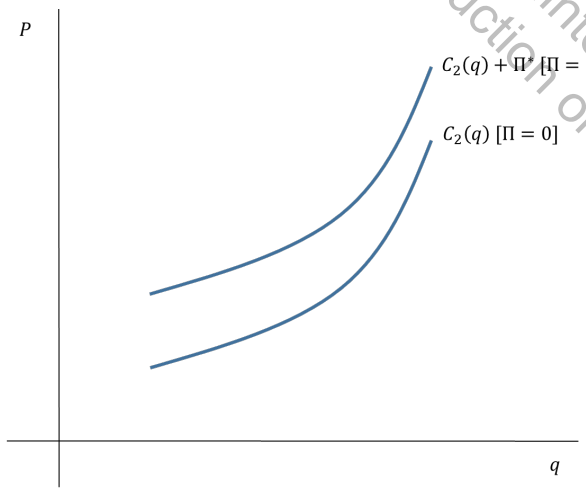


Figure 7-13: Indifference curves are just vertically shifted and thus have the same slope at every point.

Is it the case that the higher quality producers must be the ones making profit? No. The reason that this was true came from the way we drew the curves. Consider the equilibrium that would result from the situation depicted in Figure 7-14.

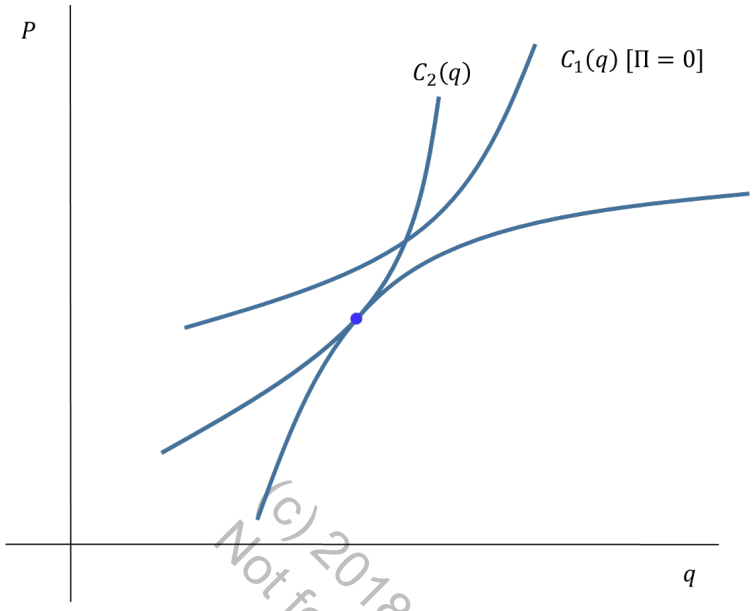


Figure 7-14: In equilibrium, the firm earning positive profits produces a lower quality good.

Heterogeneous Firms and Consumers

Now, broaden the question to multiple producer types and multiple consumer types. We will continue with the discrete case. Think about consumers A, B and firms 1, 2. B prefers high quality more than A (i.e. B's indifference curves are steeper). Then assume $C_2(q) < C_1(q)$ and $C_2'(q) < C_1{}'(q)$. That is, cost curves of 2 are below and flatter than those of 1. Finally, assume $N_1 + N_2 > N_A + N_B$, i.e. the number of producers is greater than the number of consumers, $N_1 < N_A + N_B$, and $N_2 < N_A + N_B$.

This tells us a lot about the equilibrium right away. We know $\Pi_1 = 0$, $\Pi_2 > 0$, because type 2 producers have lower cost curves. Further, type 2 produces higher quality because their marginal cost is lower, and B's want to buy from 2. Who buys from type 1? Some A's at least.

Only one type-1 iso-profit curve is relevant for drawing the equilibrium ($\Pi_1 = 0$), so we begin by drawing that one. The type-A indifference curve must be tangent to that iso-profit curve at the equilibrium quality purchased by those consumers (Q_{1A}), otherwise that quality would not be profit maximizing. See the lower-left point in Figure 7-15.

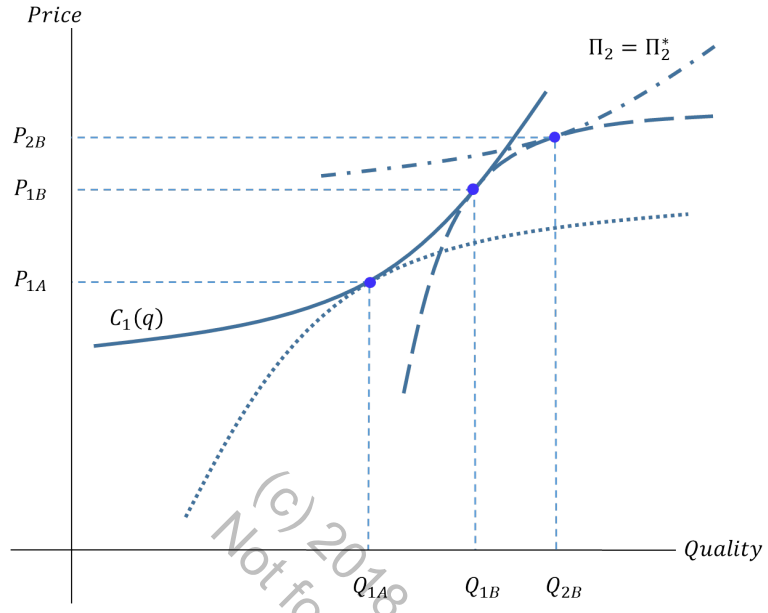


Figure 7-15: We've assumed there are more B type consumers than there are type 2 firms. So all type B consumers buy from type 2 firms (the topmost point), but then remaining type B consumers buy from the type 1 firms (the middle point). Type A consumers buy from type 1 firms.

Next we check whether any single type of firm sells to both types of consumers. Reaching a conclusion here requires an additional assumption, which we take to be $N_B > N_2$. That is, there are more B consumers than the type-2 firms can supply. Then only type B consumers buy from firm 2 because they have the stronger preference for quality. With some type-B consumers also buying from firm 1, equilibrium requires that the two transactions be on the same type-B indifference curve, which is tangent to both the type-1 iso-profit curve and the type-2 iso-profit curve at Q_{1B} and Q_{2B}, respectively. This is the equilibrium depicted in Figure 7-15.

What if we change this last assumption, and instead we have $N_B < N_2$? Now all B consumers buy from firm 2, but some A consumers will also buy from them. This equilibrium therefore has less profit than shown in Figure 7-15 (that is, the equilibrium iso-profit curve is below what is shown in the figure).

What if $N_1 + N_2 < N_A + N_B$? Now, some consumers do not get served. Boundary conditions will be determined by the utility side rather than the cost side. We'll now need a baseline utility figure to determine what level of utility is received by consumers who do not buy anything.

Note that when multiple consumers buy from the same producer, we follow the producer's cost curve, and we follow the consumer's indifference curve when the same consumer must be indifferent about buying from different producers. With a continuum of types on both sides, you get a more general picture, depicted in Figure 7-16.

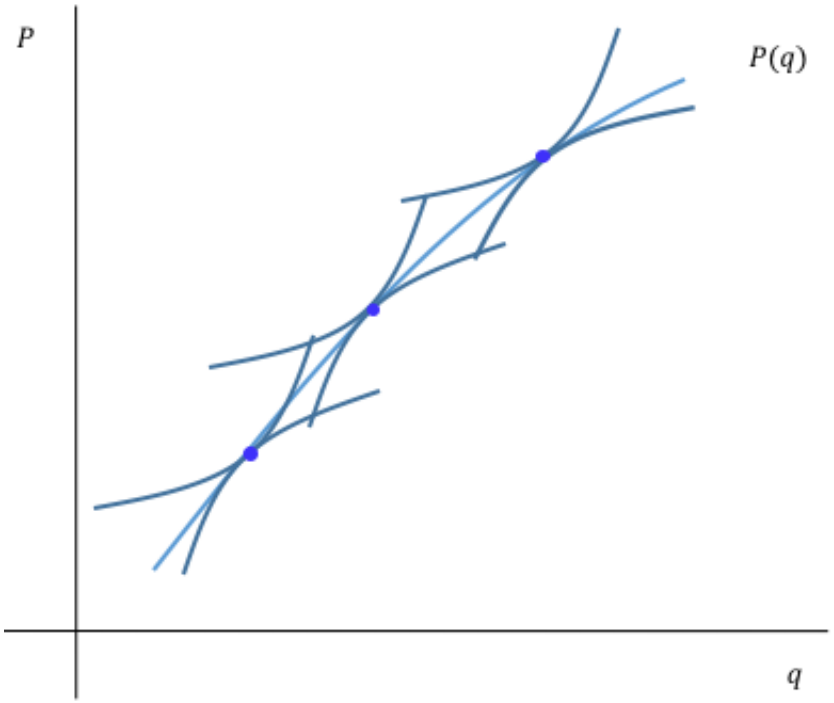


Figure 7-16: Producers are tangent to consumers, and we can trace out a function $P(q)$ from the tangencies.

Chapter 8 Location Choice: An Introduction to Equilibrium Compensating Differences

Now, think about a model of travel time. Let t be the travel time for living at t, and $R(t)$ be the rent for living at t. In equilibrium $R'(t) < 0$, that is, rents become cheaper as distance from the city increases because everyone prefers less travel time. We will not want to put travel time in the utility function, however. People do not directly care about how far they have to travel; they care about how much travel will detract from leisure time. We'll consider a simple utility function $U(C, L)$ and budget constraint $C = (24 - L - t)w - R(t)$. We can write the Lagrangian

$$\mathcal{L} = U(C, L) + \lambda[(24 - L - t)w - R(t) - C(t)]$$

This yields the first-order conditions $\frac{\partial U}{\partial C} = \lambda$, $\frac{\partial U}{\partial L} = \lambda w$, and $-R'(t) = w$. So the optimal choice of where to live has savings in rent from living another hour further away that is equal to the wage rate. Thus, we see that we've made the restriction that travel is really the same thing as work: should one work by driving one's car, or should one work at the workplace? It seems odd that being in the car would produce anything, but it does through the market: living further allows someone else to live closer to the workplace and potentially spend more time there.

Now we'll want to consider what the rent curve looks like in this model. There will be a level of rent paid by people at the center of the city $R(0)$. The slope at that point will be the wage of the highest wage worker, $-w_{MAX}$. Thus, the curve is not only downward sloping; it will be convex, because people's wages are decreasing as we move down the curve.

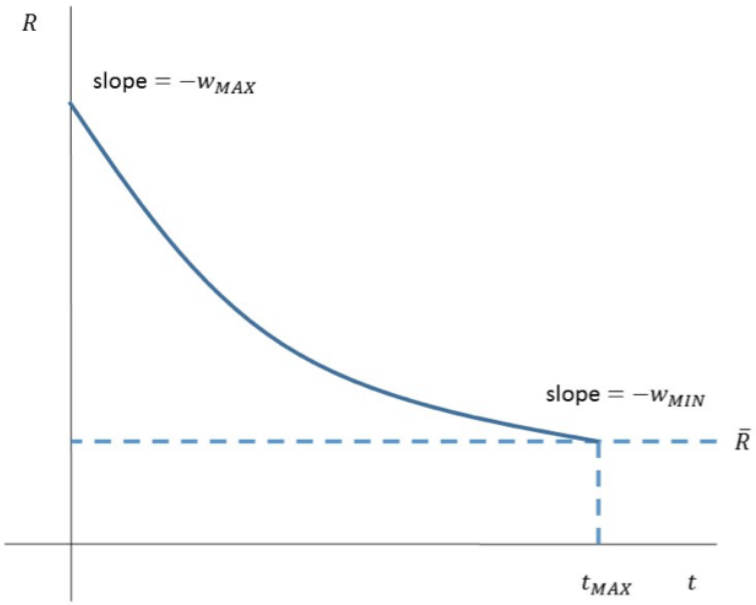


Figure 8-1: The rent gradient. $\bar{R}$ denotes what can be earned from using the land from agriculture, for example, rather than housing. In other words, $R(t_{MAX}) = \bar{R}$.

Holding L constant, we can also draw a person's indifference curves in Figure 8-1 by using the budget constraint: $(24 - L - t)w - R$. Those consumption-constant curves would be straight lines with slope

$-w$. A w person's optimal location choice is the t where his indifference curve is tangent to the rent gradient shown in the figure.

Now, how do we solve for $R(0)$, the initial boundary condition? We know the slope at that point but not yet the level. Let the boundary of the city be the distance from city center need to fit all of the city's residents, and let t_{MAX} be the amount of travel time required to reach the city boundary. In other words, at t_{MAX}, we run out of people who are driving to the city center. At that point, there is a lowest rent that will be set, for instance, by what can be earned by using the land for agriculture. Call this rent $\bar{R}$. Thus, $R(t_{MAX}) = \bar{R}$, and then we know all the slopes. So we can follow our first-order conditions about rent toward the city center until we reach the level of rents in the city and find $R(0)$.

Note that the function of equilibrium rental prices must be continuous at $\bar{R}$. If there were a jump between $R(t_{MAX})$ and $\bar{R}$, the person living at t_{MAX} could live a nanosecond further from work and receive a discretely lower rental price. This could never be an equilibrium.

Properties of the rent gradient model

The rent gradient model is illustrative of the fact that we can get a lot out of a model where preferences are very simple and we conceptualize a consumer choice problem as a production problem (in this case, cost minimization). Determining where to live had nothing to do with preferences; we simply compared the wage cost of traveling further with the rent saving. We could do the same thing for thinking about which car to buy, where the agent trades off between space and gas mileage. Moreover, the rent gradient model illustrates how hedonic models work. That is, we had that consumers preferred less travel time and less rent to more, implying the equilibrium is going to have to be downward sloping. An individual must be compensated with lower rent for needing to travel farther, and vice versa. So everyone will choose a point where they are indifferent about moving a little bit closer. Then the rent gradient is convex, because the highest wage consumers live close to the city and the lowest wage consumers live further from the city.

But now we just have a downward sloping, convex curve. How did we pin down the exact curve? We looked at the boundary condition. We considered a world in which there is some lowest rent value $\bar{R}$ set by what the land could earn for some use other than housing. Let's consider some comparative statics. Suppose we raise the wages of the lowest income people. The slope of the rent gradient would be steeper for the lowest income people, but the slope would not be affected elsewhere. Thus, rents would be higher throughout the distribution. This result is depicted in Figure 8-2.

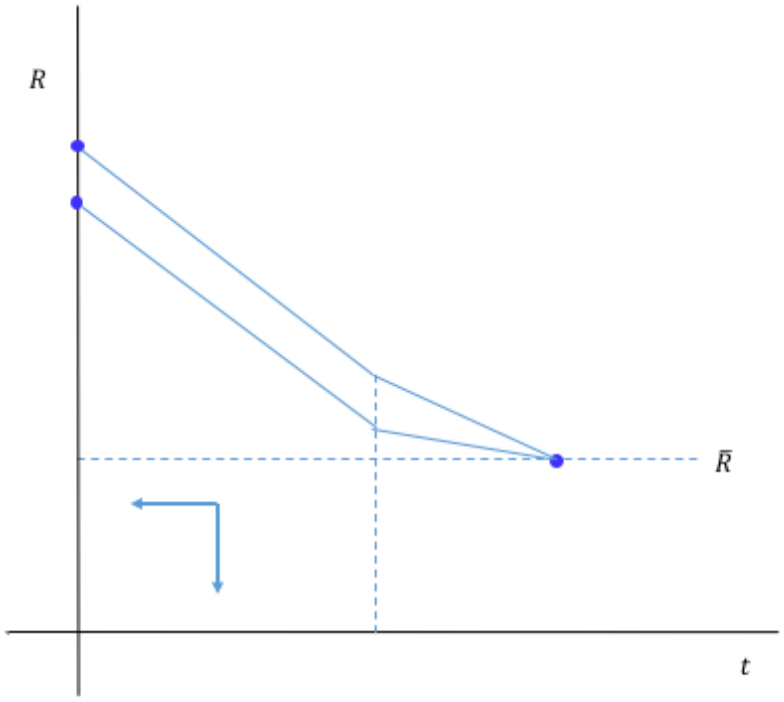


Figure 8-2: Raising wages for low income workers increases the slope of the rent gradient for wages near $\bar{R}$.

What would happen if we raised incomes of just the top half of the distribution? There would be no effect on the lower half of the income distribution, but the rent gradient for the upper half would get steeper, and rents would rise.

What would happen if we raised inequality, holding the average wage fixed? Since the average wage is fixed, the new rent gradient has the same end points. Note that a given person's rent is determined by the wages of the people who live further than he does from the city. So because their wages have gone down on average, while, for everyone above him, wages have gone up on average, it must be that his rent is lower. See Figure 8-3.

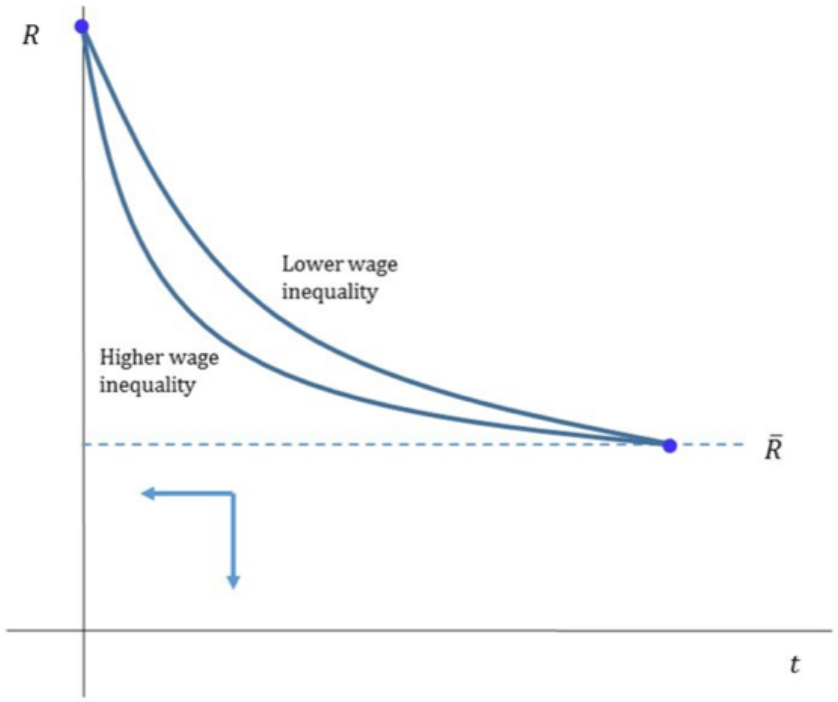


Figure 8-3: Raising wage inequality lowers rents almost everywhere (not the endpoints).

Note also that the average is weighted by distance. If roughly the same number of people live at each distance away from the city, we get the normal average. But if we think about each fixed distance from the center of the city as a circle, we might get more people living at a given distance as we move away from the city.

Consider another practical application where the locations differ in terms of their crime rate or some other housing quality metric other than distance from the city center. Let's assume poor people live in high crime areas. Let's assume we focus on people not committing crime. Why would they live in high crime areas? Low-crime areas are normal goods. Now, what if we came in and reduced crime in those areas? The people could be worse off. The rent they pay, determined by the marginal person willing to live in that neighborhood, might very well rise. And a given individual might value the reduction in crime less than the marginal people. What about a policy where we increase the quality in housing? Again, we might easily get that people are unwilling to pay the higher rents and simply move elsewhere, back into lower quality housing.

Note that the reservation rent pinned down a lot for us in this model. We could alternatively consider a world where supply of high quality housing is very elastic. Consider Figure 8-4. The elastic supply of high quality housing pins down $\bar{P}$, and prices for lower quality housing will be determined by the integral over the marginal willingness to pay of buyers for lower quality housing.

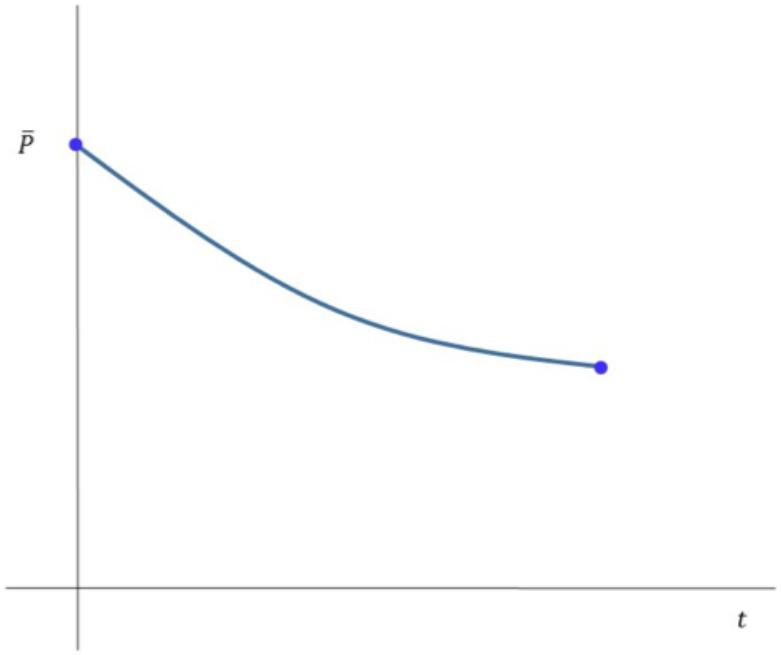
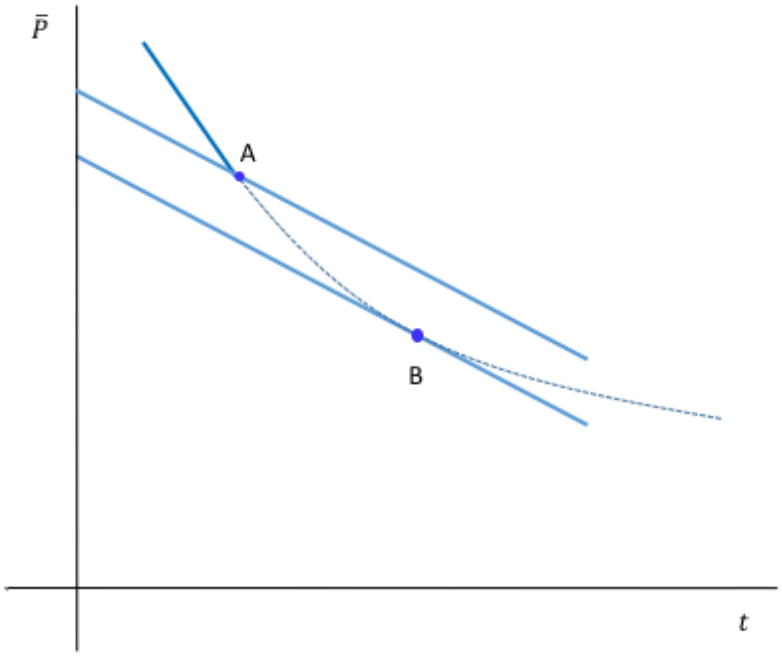


Figure 8-4: Very elastic supply of high quality housing pins down $\bar{P}$ (quality is decreasing with location along the x-axis). The price paid by someone further out with be the integral over the marginal willingness to pay of the higher quality people to his left.

As discussed previously, each person has indifference curves that can be drawn (downward sloping) in the figure. A person's optimal location – this time in terms of a crime rate – is where his indifference curve is tangent to the price gradient. For the poor person, this is probably at a point in a relatively high-crime neighborhood: to the right in the Figure.[17]

Consider the policies we considered before, where the lowest quality locations are enhanced to be the same as, say, at point A. Then those locations now have to rent for the same as point A, so policy has shifted the rent-versus-location as in Figure 7-5. The poor person may not locate at A, but he has to have utility as if he located at A because the locations to the right of A are no different. But he gets less utility at A than he was getting when his neighborhood had more crime because that extra crime allowed his rent to be lower than it is at A.

[17] Typically, buyers of lower quality housing will earn surplus, since the price is decreasing by the higher quality buyers' willingness to pay, which exceeds lower quality buyers' willingness to pay.



Lower income buyers typically do not buy low quality things because they like them. They buy them because they dislike them less than higher income people. These kinds of models help us see why that is.

Let's consider this housing model from the perspective of the model for quality that we covered earlier in the book. Suppose there are two qualities of neighborhoods, high and low. Suppose there are 60 high quality spots and 40 low quality spots for 100 total consumers. If we shift supply so that there are 70 high quality spots and 30 low quality spots, while assuming the price for high quality is pinned down (due to high elasticity of supply), then all the low quality consumers will be worse off. The reasoning is that rather than the person with the 60th highest willingness to pay being indifferent between high and low quality housing, now the person with the 70th highest willing to pay will be indifferent between high and low quality housing. To put it mathematically, initially we have $P_H - P_L = V(60)$. When we increase the number of spots in H to 70, Then $P_H - P_L = V(70)$. But P_H stays the same; then since $V(70) < V(60)$, P_L increases. People interested in living in the low-quality neighborhood are worse off.

Can we use the rent gradient model to think about urban density? Yes, if there is an increasing marginal cost for higher buildings. Your city would look like a city—that is, homes would get taller as you got closer to the city. Buildings will be of a height such that the marginal cost of an additional floor equals the value of a home at that location – which is higher closer to the city. We can also complicate the model by adding H for house size, i.e. how much house you buy. Then an individual's first-order condition becomes $w = -R'(t)H$. Whether high income people live near the city center or farther depends on whether the elasticity of demand for H with respect to income is greater than or less than 1. If that elasticity is greater than 1, people with high income want bigger houses, so they live further from the city, where land is cheaper.

Chapter 9 Learning by Doing and On-the-job Investment

In school, one learns general principles. On the job, one has to take these general principles and focus on something more specific. Someone who has just finished medical school learns how to apply these principles to patients in their residency. An electrical engineer might decide to specialize in car repairs. This is what we will call on-the-job training or investment. There are two models for this: learning-by-doing and an explicit investment model. Often, we will have a data problem. Suppose we asked how much on-the-job training there was in the US? We might look at companies. But companies frequently do not publish information about how much they are investing in their workers. Companies are not required to report this, and instead they can deduct it as labor expense. They keep accounts of capital that they must depreciate over time, but the same is not done for human capital. Often companies do not even have internal information on this.

Human capital acquired from training programs administered by the employer

If workers are paying for their training, however, how would they do it? One way would be that they are receiving lower earnings, net of their training costs. Consider Figure 9-1 below. This is the explicit investment model.

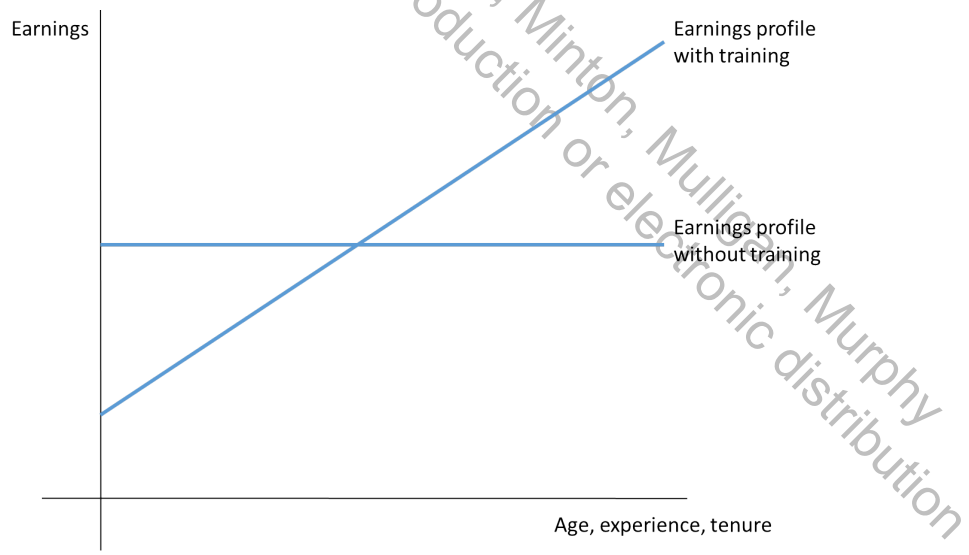


Figure 9-1: People who receive on-the-job training expect low earnings today, as the spend part of their time enrolled in training, and higher wages tomorrow when their time is more productive due to the training received.

Evaluating the human capital investment amounts to comparing the two areas in the chart: whether the earnings gains later in life justify the costs incurred early on. Alternatively, if labor-market data were showing us these two curves, we could infer how much human-capital investment people were doing. The investment amount is related to the left-hand area, although it is not identical to it.

Learning by doing

The other model is learning by doing. Learning happens as an apparently automatic byproduct of working. There is no tuition bill, and the learning does not require any time off from production or any

deliberate slowdown from production in order to learn. It is tempting to conclude from this observation that the learning is free. In other words, the wage profile might look like the upper profile in Figure 9-2 below, where the training never involves having a wage below the alternative.

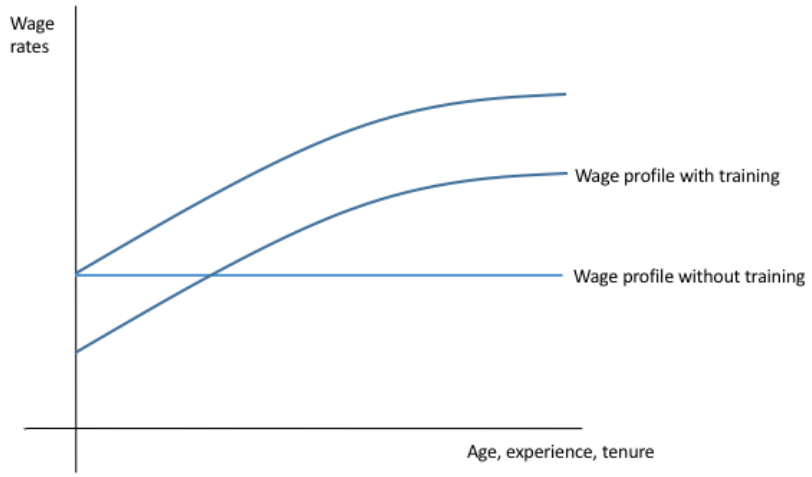


Figure 9-2: Learning by doing is not free because workers compete to get it. The wage profile (and therefore earnings profile) ends up being below the no-training alternative for a period of time, much like it is in the investment model.

This looks like a free lunch. But we know in economics there is no such thing as a free lunch. Sometimes the cost of the lunch is hidden. So where does the cost enter in here? The market equilibrium eliminates the free lunch. If the wage profile were the upper one in the figure, then everyone would want it, and no one would want the horizontal profile for the jobs without learning. The equilibrium profile for the job with learning has to be lower (see the lower profile in the figure), or the horizontal profile higher, in order for people to be willing to do both jobs. So the final learning-by-doing picture Figure 9-2 ends up looking a lot like the investment picture Figure 9-1. Indeed, Sherwin Rosen had a paper arguing that the two pictures are identical.[18] This answers one of the questions posed in the introduction of this book.

Notice that Figure 9-1 does not show wage rates, which reflect earnings per hour worked, whereas Figure 9-2 does. The training case has the added complexity as to what we mean by an hour worked: does it include time spent in the training program? If not, then the starting wage rate might not be particularly low in comparison to the starting wage rate to be earned without training. But starting earnings would be, because the training takes up time.[19] With learning by doing, there is no time spent training so this distinction does not come up. Both wage rates and earnings start out low in the learning-by-doing scenario because workers compete to be in a position that automatically gives them skill.[20]

[18] Rosen (1972).

[19] The answer might appear to be different when the training time comes out of leisure rather than work time. But this only changes the form of the cost – forgone leisure instead of forgone earnings. The amount of the cost is the same as long as leisure and work time are priced the same, as they are when time is allocated optimally (see Ghez and Becker (1975) and Heckman (1976) for models of this type).

[20] The same could be said in the context of on-the-job training if wage rate referred to the ratio of earnings to hours worked or trained.

Types of human capital

We can think about different degrees of specialization. There is firm-specific investment—investment raises productivity in one firm but not in other firms. Then there is industry-specific investment; these skills would be equally useful at any firm in a given industry. In the industry-specific case, who pays for the training? One might be tempted to say there is a positive externality because a result of a firm's investment in its worker's industry-specific skills may be that the worker applies the skills at another firm in the same industry. However, because both the firm and worker recognize this, the worker's initial wages will be lower, and the worker pays for the training.

What would be the evidence for industry-specific investment at a particular firm? If that firm goes out of business and their employees go to firms in the same industry, as Derek Neal found in his study of displaced workers (Neal, 1995), then at least part of training that was applicable elsewhere in the industry was likely acquired. How do we tell whether there was firm-specific investment? We would look at wages. If the firm goes out of business, and the employees must accept reduced wages elsewhere, then part of their human capital may have been firm-specific. Note that if the industry is small, and the company that fails is sufficiently large, there might be supply-side effects that would need to be controlled for.

Now suppose we have a monopoly, and we are thinking about industry-specific investment. The monopoly might be willing to incur these costs because it is the only firm in the industry. Note that even here, however, the worker may still pay some. To see this, suppose instead that the monopolist paid for the entire investment, and now the worker is trained and productive at the firm. The worker threatens to leave unless he is given a raise. It is a largely empty threat, because those skills are not useful elsewhere. But still there may be some bargaining and the trained worker ends up getting paid more. But then the untrained workers anticipate getting paid more after the training and compete to have the position – as in the learning-by-doing case the equilibrium result is lower earnings during the training phase.

Now we can think about firm-specific investment more generally. How does the worker know whether firm-specific investment will yield higher wages at the company later on? If the firm has done this for its skilled workers in the past, this would be a helpful signal. But in general we expect bargaining between the firm and the worker over this.

We can also look at turnover. Turnover of workers is very heavy early on and much less later on. This is similar for industry turnover. With firm-specific investment, we often think it is efficient that the workers skilled at a firm stay at that firm; turnover should be low. Separation between the employer and the worker occurs when the relationship is no longer efficient.[21]

The firm- and industry-specific distinctions are also useful for thinking about other factor markets, such as markets for raw materials, or intermediate inputs. If a user of the factors goes out of business and the factor suppliers switch to supplying someone else in the industry, then these suppliers had acquired some about of industry-specific capital.

[21] Becker did not find it useful to distinguish among various types of separation such quits, layoffs, and retirements.

Chapter 10 Production, Profits, and Factor Demand

Comparative Advantage and the Production-possibility Frontier

Now we turn to a treatment of production. We will consider Robinson Crusoe, who lives on an island and has two plots of land. On Plot A, Crusoe can produce 10 bananas or 5 oranges (or some mixture of the two). On Plot B, he can produce 15 bananas or 40 oranges (or some mixture of the two). What is Crusoe's production possibility frontier? See Figure 10-1.

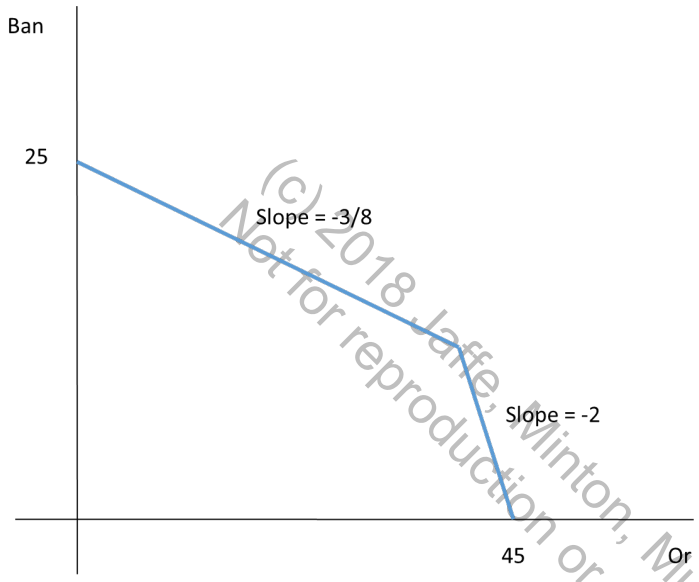


Figure 10-1: Production possibility frontier. Different slopes reflect cost differences across plots.

On Plot A, an orange costs 2 bananas. A banana costs ½ of an orange. On Plot B, an orange costs 3/8 of a banana. A banana costs 8/3 of an orange. Plot A is the low cost producer of bananas, and plot B is the low cost producer of oranges. This is what an economist calls comparative advantage. We don't care how good in absolute terms the plots are at producing bananas and oranges. A producer could be the low cost producer because it is really good at producing a particular item or really bad at producing other items.

So we get a convex production possibility frontier. The marginal cost of producing oranges rises as we produce more oranges. Why is it rising in this model? Two features are important. First, there is heterogeneity in plots, and, second, we use the lowest cost methods first. This is why we frequently assume increasing marginal cost of production. As we use more and more resources for production, we are forced to use resources that have less comparative advantage.

Note that Plot B can produce more bananas than Plot A can. Nevertheless, it is a waste for Plot B to produce *any* bananas unless people want more than bananas that Plot A can possibly produce.

For the same reason, we're not talking about the price of the plots. The price will be based in part on absolute productivity. Really bad land might be cheap to buy, but it can still have a comparative advantage over expensive plots of land.

With three plots of land, note that we get another kink. As we get lots of plots of land, we will get a more smooth convex shape, as in Figure 10-2. And the more heterogeneous the plots of land, the more convex

the shape. If Robinson Crusoe is by himself, he will find the tangency of the production possibility frontier with his indifference curve. The slope of the tangency will give the equilibrium price.

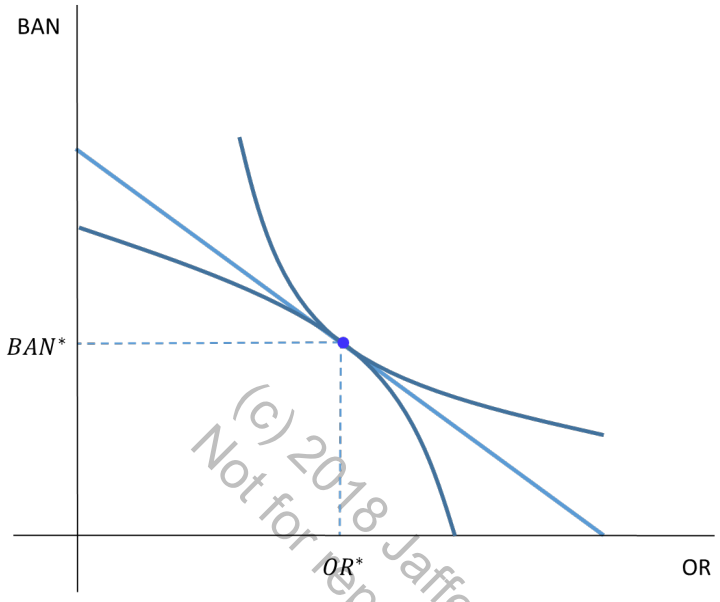


Figure 10-2: Crusoe will produce where the production possibility frontier is tangent to his indifference curve. Marginal cost will be equal to marginal value.

Now let's suppose Crusoe is allowed to join NAFTA (i.e. an agreement to trade in a world market). They will tell Crusoe the price of oranges in terms of bananas (or the reverse) in the world market. He wants this price to be *very* different from the price set from his tangency. If he gets the same price as his current price, he is no better off. Let's suppose that NAFTA tells Crusoe that oranges are more expensive in the world than on his island. This sets a trade price line with slope $-P_{OR}^{NAFTA}$, depicted in Figure 10-3. In this setting, Crusoe will produce more oranges and less bananas $(OR_{PROD}^*, BAN_{PROD}^*)$ and then trade to get $(OR_{CONS}^*, BAN_{CONS}^*)$, which is on a higher indifference curve. (Similarly, if NAFTA says bananas are more expensive, Crusoe will produce more bananas and trade a bunch of them for oranges; he will still be better off.)

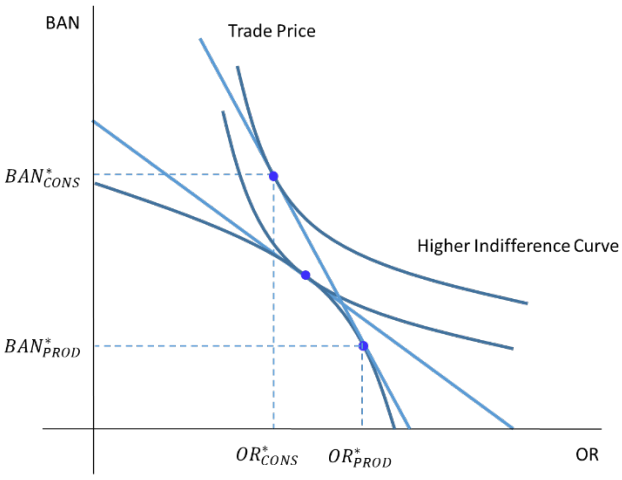


Figure 10-3: A large difference between the trade price and the price Crusoe would face without trade makes Crusoe much better off with trade.

So, what do we get out this model?

1. Trade is good. Robinson Crusoe is better off when part of NAFTA.
2. Price difference implies gains from trade. The more Robinson Crusoe's price differs from NAFTA's price, the better off he will be.
3. A theory of the firm. The firm owns a production process and chooses $(OR^*_{PROD}, BAN^*_{PROD})$. What allows this to be a theory is that we know the firm will pick $(OR^*_{PROD}, BAN^*_{PROD})$ regardless of the owner's personal preferences.
4. A competitive firm produces where price equals marginal cost. Note that this becomes an inequality when we have corner cases. If the price line is steeper than the production possibility frontier for any tangent line along the frontier, then Crusoe will produce only oranges and, further, we will not have that marginal cost equals marginal value.
5. Marginal cost = marginal value. Again, this may not hold in corner cases.

This production example has a lot in common with Chapter 7's result that the poor could be harmed when their neighborhood gets safer. In both examples, the benefit from trade comes from being able to make different choices from everyone else. Crusoe produces oranges and trades them for the bananas (produced by others) that he wants. A poor person may take the low-quality housing in order to get more of the other goods that he wants while in the same market richer people give up other goods in order to get the high-quality housing that they want. Trade makes these choices possible.

The Production Function

Now we will want to consider how a firm behaves. Consider the following production problem. We will have a production function, output Y, and inputs $X_1, \ldots, X_N$, and we will specify

$$Y = F(X_1, \ldots, X_N)$$

The production function is the result of maximization already – F returns the maximum amount of output achievable using the inputs. Just going out and purchasing the inputs by itself will not make the output magically appear. The inputs need to be put together in the right way. The details of managing and combining the inputs can be interesting, but whenever we start with a production function we have put those details aside. Nevertheless, you will see that we have some interesting things to say about how firms interact with the rest of the marketplace.[22]

The production function is said to exhibit constant returns to scale if the function is homogeneous of degree one in the inputs. Equivalently, if we calculated the elasticity of output with respect to each input, those elasticities would sum to one. If the sum is less (greater) than one, then the production function is said to exhibit decreasing (increasing) returns to scale, respectively. Decreasing returns to scale is a common assumption to make when looking at a particular firm.

[22] Gary Becker took the same approach to analyzing the public sector (Becker 1983, Becker 1985, and Becker and Mulligan 2003): posit a public-policy production function that subsumes the details of political processes but nevertheless has interesting things to say about how public policy making interacts with the rest of the economy. However, Becker's approach has been resisted by much of the political economics literature where it is asserted that political details are essential for predicting policy outcomes (e.g., Myerson 1995).

We will also assume the firm is competitive. That is, the firm does not affect (1) $P =$ price of output or (2) $w_1, \dots, w_N$, the prices of inputs. Is there a difference already between this and utility? We can measure the output here, whereas we couldn't do this with utility.

Profit Maximization

Firms will maximize profits. Why is this a reasonable assumption? In consumer theory we say that more income permits an individual to obtain a higher utility level. An individual owning a firm therefore gets more utility the more profit income that he obtains from that firm. If, on the other hand, the activities of the firm directly enters the person's utility function – having lots of capital, for example – this will not be a good assumption.[23] Owners of basketball teams typically care about winning; it is not necessarily a priority for them to have the most amount of money possible at the end of the season. Gary Becker's dissertation was about the idea that some firms like to hire some people more than others—for example, they might want to discriminate.[24] So this assumption may miss some of what's going on, but it is typically a good first-order approximation to think about firms as maximizing profits. We will denote profits as revenue minus cost, $PY - \sum_{i=1}^{N} X_i w_i$.

$$\max PF(X_1, \dots, X_N) - \sum_{i=1}^{N} X_i w_i$$

This looks a lot like our utility maximization problem. But, here you can trade the output. You can't trade utility. Even if we could quantify utility, these problems would still be different. Production of utility is intrinsic to consumption. Output, on the other hand, is ordinal, measurable, and transferrable. This is why we got so much more traction with the rent-gradient model after turning it into a production problem. In that model, people ended up working where it was most productive—either in their car or at work.

The profit-maximization problem leads to the first-order conditions

$$P\frac{\partial F}{\partial X_1} - w_1 = 0$$

$$\dots$$

$$P\frac{\partial F}{\partial X_N} - w_N = 0$$

This is similar to our results from the consumer problem, but now we lack the multiplier. In the consumer problem, the (endogenous) multiplier converts dollars into utility, but since output is sold at price P, that is the value of a unit of output; the price replaces the multiplier. Rearranging, we can write the marginal product in terms of prices:

$$\frac{\partial F}{\partial X_1} = \frac{w_1}{P}$$

$$\dots$$

[23] Lakdawalla and Philipson (2006) analyze nonprofit businesses in this way.

[24] Becker (1957).

$$\frac{\partial F}{\partial X_N} = \frac{w_N}{P}$$

Further, we can solve this system of equations to get input-demand functions

$$X_1 = X_1(w_1, \dots, w_N, P)$$

...

$$X_N = X_N(w_1, \dots, w_N, P)$$

We can also derive a resulting supply function:

$$Y = Y(w_1, \dots, w_N, P)$$

Recall that before we wrote $P = MC$. The system of first-order conditions on the previous page is identical to setting price equal to marginal cost. Just rearrange the first-order condition to get that $\frac{w_i}{\frac{\partial F}{\partial x_i}} = P$. This is saying that dollar cost per marginal unit of output for increasing that input equals the output price. See also Figure 10-4.

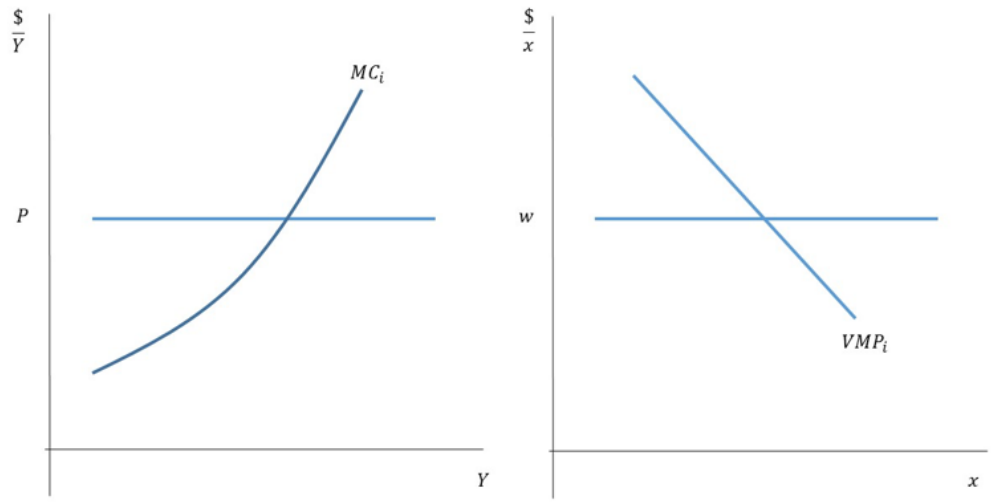


Figure 10-4: Setting wage equal to the value marginal product is the same as setting price equal to marginal cost.

Asking "How much input should I use?" is identical to asking "How much output should I produce?"—we're simply using a different perspective.

Cost Minimization

The new problem has two steps. In step 1, minimize cost for a given level of output. In step 2, pick the level of output to maximize profits.

STEP 1

$$\min \sum X_i w_i \qquad s.t.\ F(X_1, \dots, X_N) = Y.$$

This problem, in turn, yields the cost function $C(w_1, \ldots, w_N, Y)$. This is like solving for Hicksian demand functions. Now the constraint, instead of being a utility function, is the production function. Importantly, this time we can tell empirically if a firm is holding output constant. As with the expenditure function, taking the derivative of the cost function with respect to the wage yields demand:

$$\frac{\partial C}{\partial w_i} = X_i(w_1, \ldots, w_N, Y)$$

These are called *conditional factor demands* because they are conditional on output. This is the end of step 1.

STEP 2

$$\max PY - C(w_1, \ldots, w_N, Y)$$

Sometimes the problem will start here. Just remember that this function C implicitly assumes a production function. The first-order condition is

$$P - \frac{\partial C}{\partial Y} = 0$$

which is again the fact that $P = MC$.

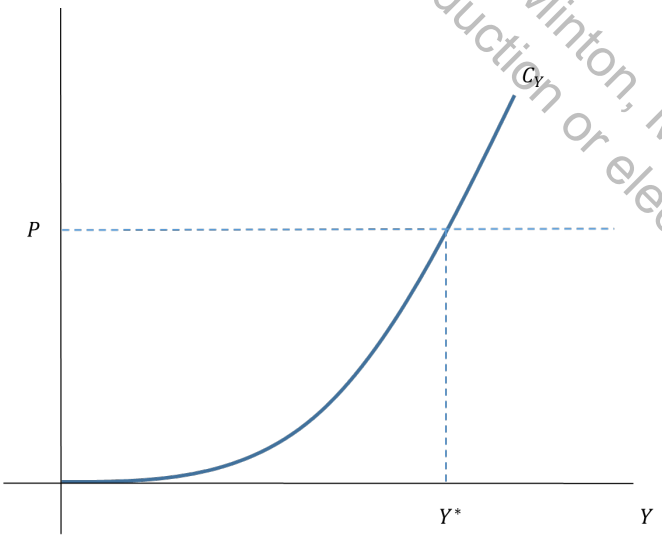


Figure 10-5: The firm chooses the level of output so that price equals marginal cost.

Recall in the consumer problem we could go back and forth between the Marshallian and Hicksian demand curves using the Slutsky equation. Unconditional factor demands are different from Marshallian demand because firms can purchase more input by being efficient producers, but individuals cannot purchase more goods by being efficient producers of utility.

As an aside, think about why this problem behaves well. For instance, why is marginal cost increasing? Equivalently, why does $F(X_1, \ldots, X_N)$ exhibit decreasing returns to scale?[25] Why can't a firm just replicate current inputs to double output? It might be difficult to replicate exactly what exists. No set of workers for hire may be exactly the same as currently hired workers. Even if such workers could be found, it might be hard to find them. Similarly, even if the same inputs can be found, some amount of time is required for replication. Once double the operation is occurring, firms also tend to require additional administration, raising marginal cost.

A common practice in applied industrial organization is to look at the expenses of a business and declare that materials and production employees are marginal costs but that management, space, and other "fixed" expenses are not marginal costs. Naturally these studies find that the business charges more than the narrowly-defined marginal costs, but this may not tell us that price exceeds marginal cost. Rather, the business recognizes that engaging in more production will hasten the date, or increase the probability, that management, space, and other so-called fixed expenses have to be expanded. These are marginal costs too.

The Firm's Slutsky Equation

Though we do not have the Slutsky equation, there is an equation relating unconditional and conditional factor demands. Recall, $\frac{\partial C(w_1,\ldots,w_N,Y)}{\partial w_i} = X_i(w_1, \ldots, w_N, Y)$, and $P = \frac{\partial C(w_1,\ldots,w_N,Y)}{\partial Y}$. Totally differentiate with respect to w_1, allowing Y to adjust (that's what we mean by unconditional factor demand), to get

$$\frac{\partial X_i}{\partial w_i} = \frac{\partial^2 C}{\partial w_i^2} + \frac{\partial^2 C}{\partial w_i \partial Y}\frac{dY}{dw_i}$$

$$0 = \frac{\partial^2 C}{\partial w_i \partial Y} + \frac{\partial^2 C}{\partial Y^2}\frac{dY}{dw_i}$$

Eliminating $\frac{dY}{dw_i}$ yields

$$\frac{\partial X_i}{\partial w_i} = \frac{\partial^2 C}{\partial w_i^2} - \left(\frac{\partial^2 C}{\partial w_i \partial Y}\right)^2 \Big/ \frac{\partial^2 C}{\partial Y^2}$$

This is the firm's version of the Slutsky equation. We know that $\frac{\partial^2 C}{\partial w_i^2} \leq 0$ because the cost function is concave in prices. We also know that $\frac{\partial^2 C}{\partial Y^2} > 0$, since this is the change in marginal cost as we increase output. If marginal cost is falling, we would be finding a minimum, so it must be rising. Now the squared term cannot be negative, so we can conclude that X_i is not increasing in w_i. Unlike in the consumer problem, where we have Giffen goods, input demand cannot increase in its factor price.

Now what can we say about output? The sign is actually ambiguous. The firm may increase output as w_i increases if it can better avoid higher cost by increasing output. If two inputs are substitutable, for example, and the less productive input becomes more expensive, the firm may increase total output when it shifts to using the more productive input.

[25] With constant returns to scale, marginal cost is independent of output and the cost function is proportional to output.

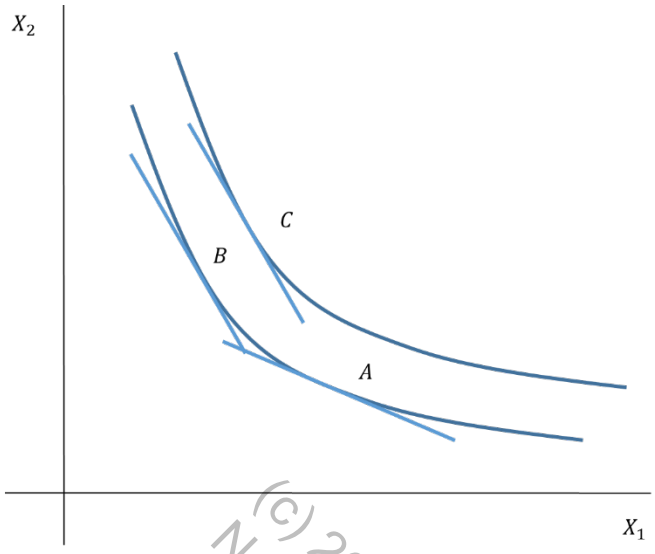


Figure 10-6: Consider an increase in the price of X_1. The substitution effect leads the firm to shift from A to B. Though total costs have gone up, in this case marginal cost went down so output is increased—the scale effect leads the firm to shift from B to C.

For a given level of output, total (and average) cost has still gone up, but the marginal cost of output may have decreased, leading to an increase in output. However, the case of a factor price reducing marginal cost is rare. The factors associated with these cases are called *inferior factors*.

Two-Input Production

Often, we write output as a function of capital and labor:

$$Y = F(K, L)$$

K has a capital or purchase price and a rental price, R. The capital price measures the value over the life of the unit of capital. The rental price, on the other hand, measures the price of current use. For example, renting computers that would cost $1 million to buy is much more expensive than renting factories that would cost $1 million to buy, because the factories will produce its $1 million in value over decades, whereas computers become obsolete quickly and must produce their value quickly.

Labor has the rental price w. Paying the wage gives the firm access to a person's labor for a specified amount of time. So firms will solve

$$\max PF(K, L) - wL - RK$$

In this case, we get the same first-order conditions as before:

$$P\frac{\partial F}{\partial L} = w$$

$$P\frac{\partial F}{\partial K} = R$$

One useful result from these conditions is that we might measure marginal cost as the ratio of either factor's rental price to its marginal product – either ratio should give the same answer.

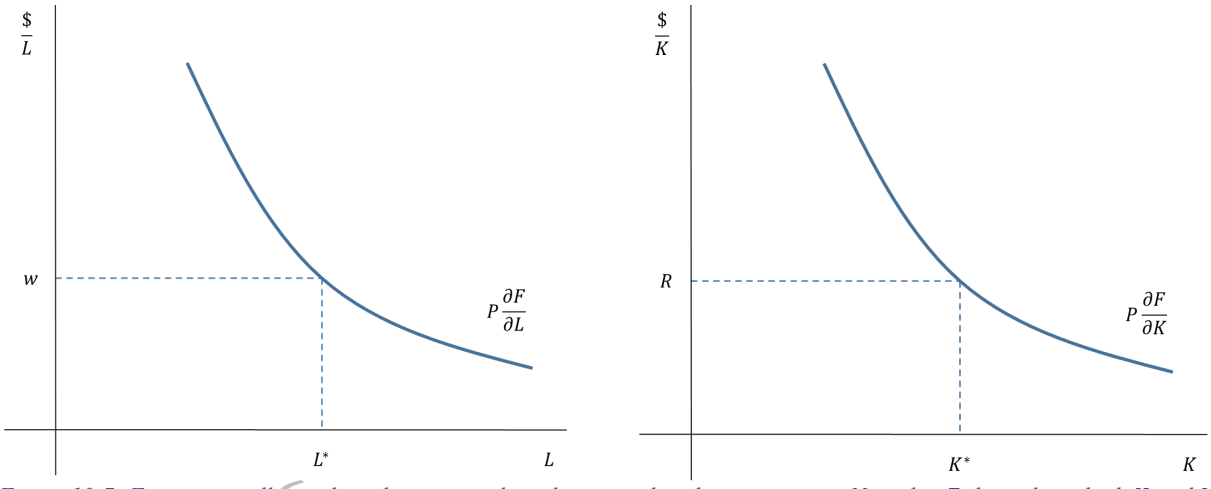


Figure 10-7: Firms optimally set the value marginal products equal to the input prices. Note that F depends on both K and L, so its derivative might as well. That is, the value marginal product curve in the left graph may change as K changes, and similarly in the right graph.

When thinking about capital, it is important to think about the short run versus the long run. A common view in production is that capital is fixed in the short run but variable in the long run. Labor has traditionally been viewed as much more flexible. These notions have become less useful over time. A law firm trying to expand, for example, will have a much harder time building an integrated labor force than finding office space, computers, etc.

Now, consider a reduction in the wage rate, from w to $\hat{w}$. This shift is depicted in Figure 10-8. In the short run, labor will increase. In the long run, we will need to think about the cross partial derivatives of the production function—whether labor and capital are complements or substitutes. Suppose that $F_{LK} > 0$, i.e. that the marginal productivity of labor is increasing in the amount of capital. Then since we have increased labor in the short run, we will optimally increase capital over time. This will have additional feedback effects with labor, since increasing capital will further make labor more productive. Eventually, we approach a long-run equilibrium with a higher level of labor and capital.

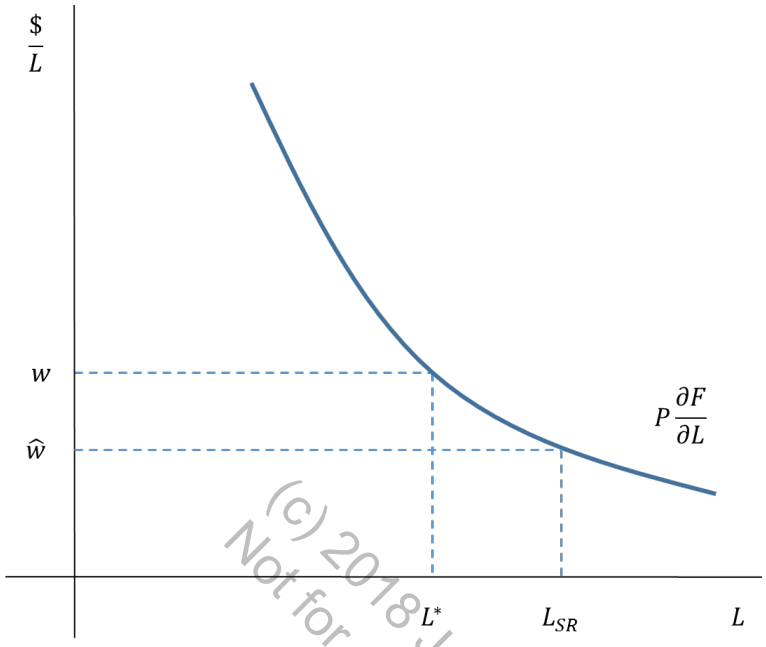


Figure 10-8: A decrease in the wage increases labor in the short run.

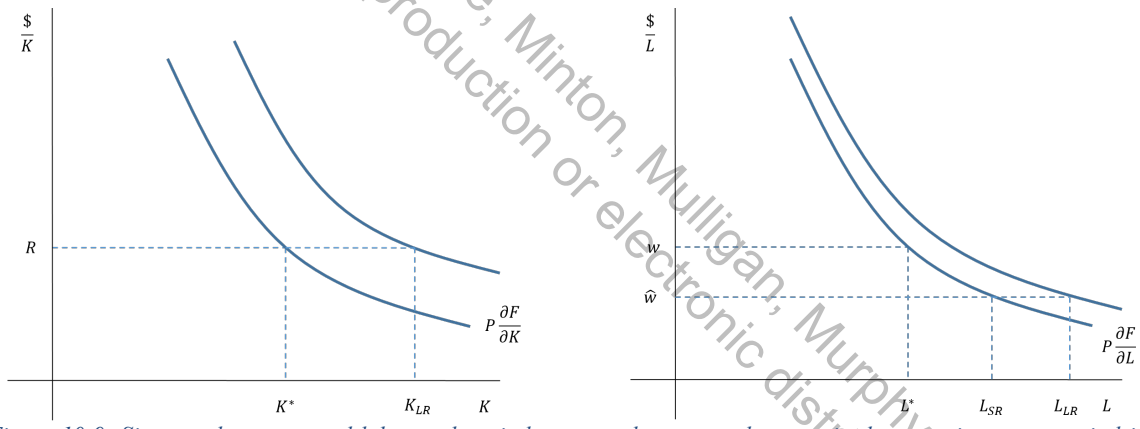


Figure 10-9: Since we have assumed labor and capital are complements, a decrease in the wage increases capital in the long run. As the firm increases capital, it will also seek to increase labor further. Thus, the economy eventually approaches the depicted long run equilibrium.

Now, what about the case where $F_{LK} < 0$? This means that capital and labor are substitutes. In this case, the firm would choose optimally to *reduce* capital, since it can substitute the cheaper input of labor. The feedback effects still exist, but they serve to drive capital lower and increase labor more. Again, labor is more elastic in the long run than in the short run. Since capital is a complement to labor if and only if labor is a complement to capital, the long run effect always magnifies the short run effect.

The same is often, though not always, true for consumption. If the price of gas goes up, in the short run people will drive less. If cars and gasoline are complements, in the long run they will buy fewer cars and drive even less. If biking and driving are substitutes then in the short run people will bike more and in the long run more people will buy bikes and biking will increase even more. Either way the long-run effect is larger than the short run effect.

This idea of feedback effects, however, deserves some mathematical discussion. These effects converge because of the second order condition of the problem. We have assumed decreasing marginal

productivity, so, for a generic objective function $f(X,Z)$, we know that $f_{XX} < 0$ and $f_{ZZ} < 0$.[26] For profit maximization, we also need that $f_{XX}f_{ZZ} - f_{XZ}^2 > 0$. The simple approach to solving this problem we have discussed by mentioning feedback effects is akin to solving the maximization problem in two steps. Step 1 is to maximize over one variable (say, *Z*) for a fixed value of the other

$$G(X) = \max_Z f(X,Z)$$

At the optimal Z^*, we have $f_Z = 0$, $f_{ZZ} < 0$. Total differentiation of the equality gives $f_Z + f_X \frac{dZ}{dX} = 0$. We also have

$$G_X = \frac{df(X,Z^*(X))}{dX} = f_X + f_Z \frac{dZ^*}{dX} = f_X$$

Total differentiation gives $G_{XX} = f_{XX} + f_{XZ}\frac{dZ}{dX}$. We can plug in for $\frac{dZ}{dX}$ to achieve that

$G_{XX} = f_{XX} - \frac{f_{XZ}^2}{f_{ZZ}} < 0$.

This determinant tells us that once we optimize over one variable as a function of the second, the problem is still concave in the second variable. In step 2, maximize over the other variable, and repeat this process.

The "Slutsky" equation for the firm discussed above can also be interpreted as a sequential optimization problem: get the right factor amounts for a given output, and then get the right output. Recall that we can access the labor demand function from the cost function, in the same way that we could derive Hicksian demand functions from the expenditure function. That is, $L = \partial C(w,R,Y)/\partial w$. Totally differentiating this, allowing output to adjust with *w*, yields that $\frac{\partial L}{\partial w} = \frac{\partial^2 C}{\partial w^2}w + \frac{\partial^2 C}{\partial w \partial y}\frac{dY}{dw}$. Similarly, use the first-order condition from step 2 of the cost version of the firm problem: $P = \frac{\partial C(w,R,Y)}{\partial Y}$. Again, totally differentiating yields $0 = \frac{\partial^2 C}{\partial w \partial y} + \frac{\partial^2 C}{\partial y^2}\frac{dY}{dw}$. Combing these two results to eliminate the term $\frac{dY}{dw}$ gives

$$\frac{\partial L}{\partial w} = \frac{\partial^2 C}{\partial w^2} - \left(\frac{\partial^2 C}{\partial w \partial y}\right)^2 \Big/ \frac{\partial^2 C}{\partial y^2}$$

Substitution and Scale Effects on Factor Demand

We have repeatedly expressed the Slutsky equation for the firm in terms of own-price. That is, we have looked at the effect on the demand for input X_i if we increase the price of X_i. In the same exercise we have done twice now, one can show more generally that

$$\frac{\partial X_i}{\partial w_j} = \frac{\partial^2 C}{\partial w_i \partial w_j} - \left(\frac{\partial^2 C}{\partial w_i \partial Y}\frac{\partial^2 C}{\partial Y \partial w_j}\right) \Big/ \frac{\partial^2 C}{\partial Y^2}$$

[26] In the video, Professor Murphy wrote *F*(*X*,*Y*) but here we replaced *F* with *f* and *Y* with *Z* because *F* and *Y* are used earlier in the lecture for different purposes. An example of a generic objective function *f* would be firm profits, part of which is the production function *F*.

The result obtained before is the special case $j = i$. Now, consider the two terms on the right hand side of the equation. The left term, $\frac{\partial^2 C}{\partial w_i \partial w_j}$, is called the *substitution effect*. The substitution effect tells us how much input *j*'s price induces a shift along the isoquant toward input *i*. Since the cost function is the result of firm optimization, it carries with it the assumptions of the underlying production function used in step 1 of the problem.

The right term, $\left(\frac{\partial^2 C}{\partial w_i \partial Y}\frac{\partial^2 C}{\partial Y \partial w_j}\right) / \frac{\partial^2 C}{\partial Y^2}$, is called the *scale effect*. It is the effect of the factor price on factor demand through changes in the scale of production. Since $\frac{\partial C}{\partial w_i}$ is the demand function X_i, $\frac{\partial^2 C}{\partial w_i \partial Y} = \frac{\partial X_i}{\partial Y}$. X_i is an *inferior input* if and only if $\frac{\partial^2 C}{\partial w_i \partial Y} < 0$. Note that a factor price cannot reduce cost but it can, and does in the inferior-input case, reduce marginal cost. Lower marginal cost means that output can expand when one of the factors gets more expensive. Maybe a firm is using shovels to do small digging projects. But then shovels get more expensive, so the firm switches to digging with an excavator machine. As long as the firm has the machine, the firm digs more.

Inferior inputs are not that common, and we now see from the firm's Slutsky equation that additional production- or cost-function restrictions are needed to guarantee that no inputs are inferior. Moreover, with *i* = *j* the Slutsky equation's scale effect term must be negative even though the direction of the scale effect is ambiguous – inputs can be normal or inferior. With a normal input, scale and factor demand move together and the factor price reduces scale. With an inferior input, scale and factor demand move in opposite directions but the factor price increases scale. Either way, the factor price is reducing factor demand through the scale effect. The scale effect of factor prices is always reinforcing the substitution effect.

Acquired Comparative Advantage

We began this chapter by assuming that production factors just happened to be different in ways that created comparative advantage. But the marketplace gives people an incentive to become different, to strengthen their comparative advantage. To explore this, we begin with the simplest example of comparative advantage and then add human capital acquisition to it.

Think about a simple world with two tasks, A and B. An individual has human capital for those tasks H_A and H_B. Whatever task he picks, he is paid a wage per unit human capital: w_A or w_B as appropriate. This will mean total income for an individual from task A is $Y_A = w_A H_A$ and from task B is $Y_B = w_B H_B$. If an individual can only choose one task, the individual will choose to earn income

$$Y = \max(w_A H_A, w_B H_B)$$

So the individual picks task A if $w_A H_A > w_B H_B \Leftrightarrow \frac{w_A}{w_B} > \frac{H_B}{H_A}$. This is comparative advantage because his choice depends on the relative amounts of human capital that he has, not the absolute amount.

We illustrate the choice in the $[H_A, H_B]$ plane by drawing a task-indifference ray showing all of the configurations of human capital that someone could have and be indifferent between the two tasks. See Figure 10-10.

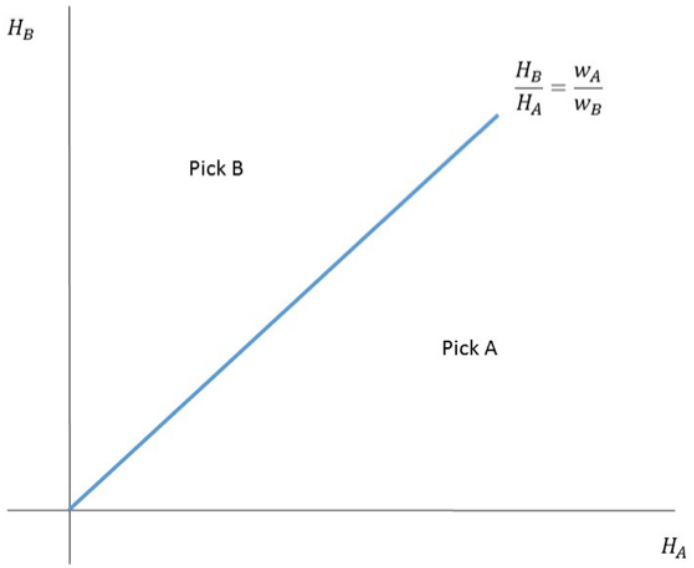


Figure 10-10: Supply and demand will rotate the task-indifference ray until the right number of workers is in each task.

There is demand for tasks A and B, which in equilibrium has to match up with the available human capital and the aforementioned incentives for workers to choose one task rather than the other. This happens with wage adjustments. If there were a lot of demand for A, then Figure 10-10's task-indifference ray has to be steep so that lots of workers choose task A and few choose B. In other words, w_A/w_B would be greater than one.

Now, assume we have reached the equilibrium, so that w_A/w_B reflects market supply and demand. Then for any point on the line, every person directly below and directly left must be earning the same income. See the dashed lines in Figure 10-11. This is because each person on the dashed line above the task-indifference ray has the same level of H_B and his H_A does not matter because he does not use it. Each person on the dashed line below the task-indifference ray has the same level of H_A and his H_B does not matter because he does not use it. Let's call the union of the two dashed lines an indifference curve for the worker.

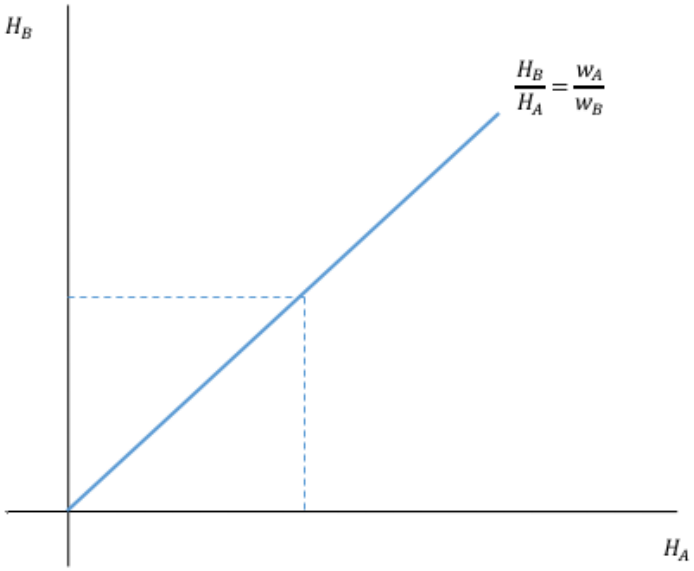


Figure 10-11: A worker indifference curve: each person on the dotted line earns the same income.

Now, let's allow each agent to choose their human capital. For example, they are considering whether to be a good plumber versus being a good carpenter. The opportunity set for human capital could have an interesting shape, as depicted in Figure 10-12. Consider the point associated with the maximum level H_B. As it is depicted, this person will have some positive level of H_A. This reflects an underlying story that some of the tasks A and B require some of the same abilities. Thus, if the agent chooses to be a good plumber, that doesn't mean that the agent ends up with zero human capital as a carpenter.

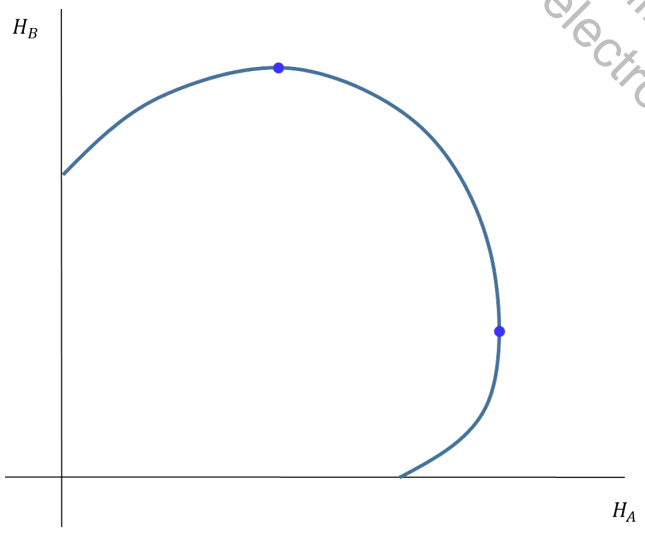


Figure 10-12: The opportunity set for selecting human capital. The agent with maximum human capital for task A still has positive human capital for task B.

Note further, in this graph, that the economically relevant region of the opportunity set lies between the two points, and we can erase the parts of the curve close to the axes because no one would chose a human capital pairing left of the top point or below the right point. On the erased regions, the agent could be better at both tasks!

Now let's put the opportunity set together with the worker's indifference curves, as in Figure 10-13. We can even have everyone identical in the sense that they all have the same opportunity curve to choose from. Nevertheless, specialization is optimal behavior. Being equally good at tasks *A* and *B* is worse than being very good at just one task because you have acquired a lot of human capital that you do not use.

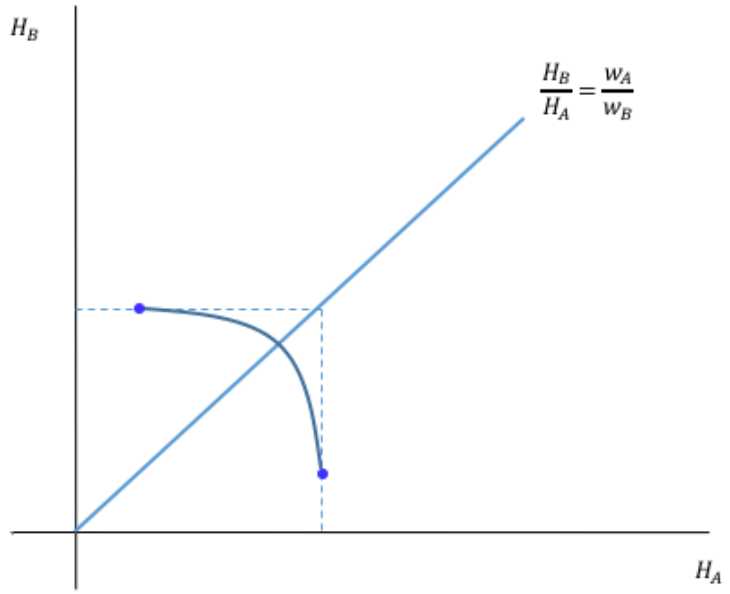


Figure 10-13: Specialization. Agents maximize their human capital at task A or task B.

We started this picture by indicating the types of workers (that is, configurations of human capital) who are indifferent between the two tasks. But now we have shown that people will not choose to be those types of workers. Because the human capital is acquired, it is not an equilibrium for people to be indifferent between the two tasks.[27]

The equilibrium requires that both tasks are performed, so some people specialize in *A* and the others specialize in *B*. People who are identical in the sense of having the same opportunities open to them end up being different.

One might say that it is a coin flip exactly who goes toward task *A* and who goes toward task *B*, and we would agree if people were precisely identical. But in reality, people have somewhat different opportunities open to them: in Figure 10-13, that means somewhat different opportunity curves. Some of the opportunity curves may be relatively steep and others relatively flat. Then just a small difference among people in the slope of the curve will decide who specializes in what. Specialization in the marketplace turns small differences into large differences.

Gary Becker revolutionized labor economics by showing how so many of the differences among workers are acquired.[28] They did not just happen independent of supply and demand considerations.

[27] This simple model abstracts from timing, uncertainty and other factors. In the more general case, the market may induce some people to be on the task-indifference ray because, at the time that they acquire skills, they do not know which task they will end up doing. But even in this case, it will not make sense for everyone to be near that ray: some of them can be confident that they will be doing a particular task and thereby specialize in it.
[28] Becker (1964) and Becker and Murphy (1992).

Chapter 11 The Industry Model

Properties of the Industry Model

We begin with the industry model of demand and assume constant returns to scale at the industry level.[29] Constant returns to scale is a somewhat problematic assumption at the firm level, because if the price is above cost, it is optimal to produce an infinite amount of output; if the price is below cost, it is optimal not to produce; and if the price is equal to cost, output is indeterminate. Since firm-level equilibria are often indeterminate it is often necessary to move to the industry level to begin considering equilibria. When we assume the industry exhibits constant returns to scale, we are *not* necessarily assuming that firms have constant returns to scale. Each firm in the industry could have diminishing returns to scale, but expanding the industry as a whole can exhibit constant returns to scale—replication may be feasible on the industry level. So assuming constant returns at the industry level encapsulates two cases: (1) firms in the industry exhibit constant returns to scale, or (2) firms in the industry exhibit diminishing returns, which in the aggregate are well-approximated by a constant returns to scale model of the industry.

We further assume the two-input case. Then constant returns to scale means

$$F(tL, tK) = tF(L, K)$$

For our conditional factor demand functions, constant returns to scale will imply

$$L^*(w, r, Y) = YL^*(w, r, 1)$$

$$K^*(w, r, Y) = YK^*(w, r, 1)$$

The derivatives are therefore homogenous of degree zero:

$$\frac{\partial F(tL, tK)}{\partial L} = \frac{\partial F(L, K)}{\partial L}$$

$$\frac{\partial F(tL, tK)}{\partial K} = \frac{\partial F(L, K)}{\partial K}$$

[29] Recall from Chapter 7 that industry demand is the sum of each consumer's demand.

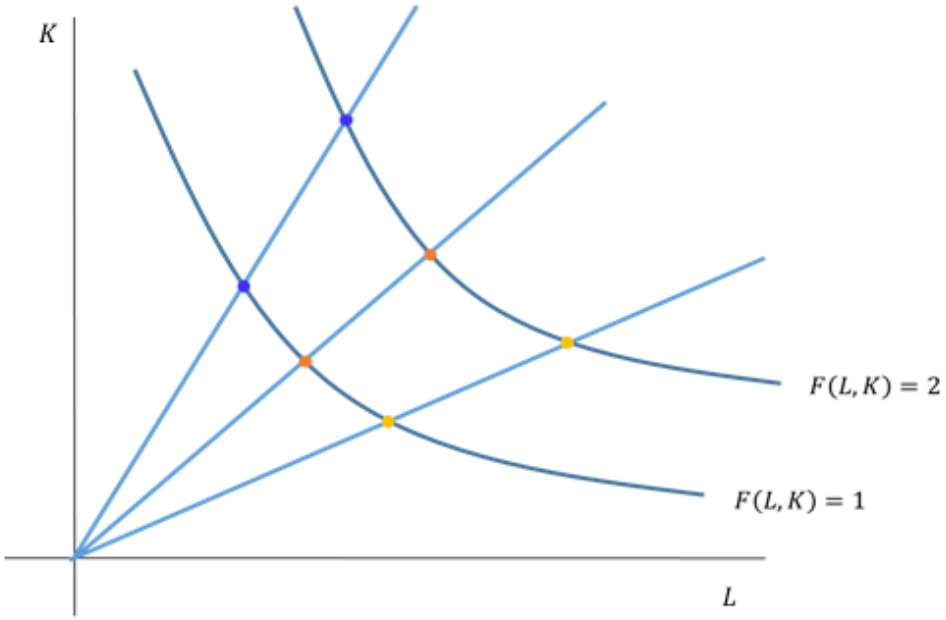


Figure 11-1: Doubling inputs is associated with doubling output. Further, at each of the same colored points, the slopes of the isoquants are the same.

We can also make statements about the cost function:

$$C(w,r,Y) = YC(w,r,1)$$

For the cost function, constant returns to scale implies that the cost of producing Y units is the same as the cost of producing 1 unit Y times. Taking the derivative, $\frac{\partial C(w,r,Y)}{\partial Y} = C(w,r,1)$, so that marginal cost is constant. In sum, constant returns to scale says that relative factor prices determine the ratio of K to L, and the amount of output will determine the levels of K and L required. So CRS gives a convenient decomposition of relative input use and the level of output.

It remains to determine how optimal output is pinned down. Recall the condition $P = MC$. In this context, that means $P = \frac{\partial C(w,r,Y)}{\partial Y} = C(w,r,1)$. We also have market clearing, $Y = D(P)$, so that the amount of output equals the amount of the good consumers demand.

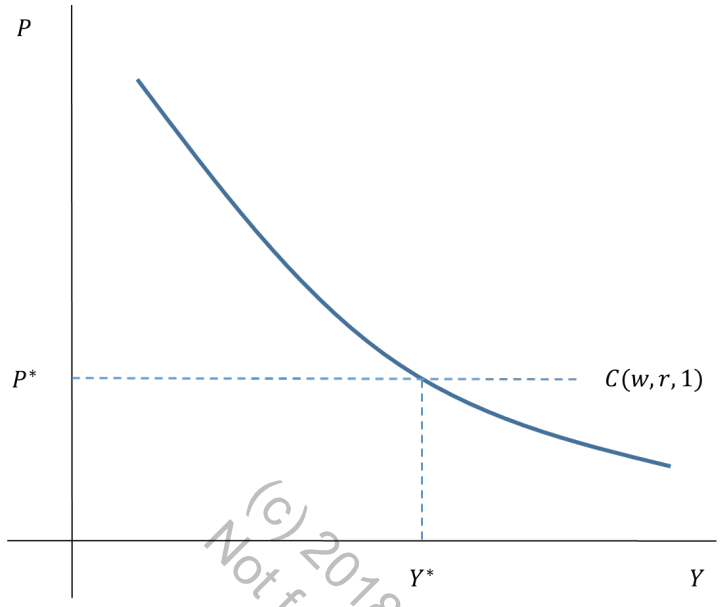


Figure 11-2: Optimal output is determined by the intersection of the constant marginal cost curve, which equals the price, and consumer demand.

Supply is perfectly elastic, so the supply side determines the price. The quantity is then determined by how many consumers are willing to purchase the good at the determined price.

As an aside, note that for most industries, this is a good way to think about the world. Prices are determined on the supply side by what it costs to produce the product. The quantity we consume is determined on the demand side by how many goods we want to consume at the given price.

The Supply-Demand Perspective on Industry Behavior

The supply and demand perspective has a lot to say about important issues ranging from inequality (Katz and Murphy 1992) to the war on drugs (Chapter 12 of this book) to the business cycle (Mulligan 2012). Although it is something that should be mastered in undergraduate studies, even the experts sometimes get confused about what the supply and demand perspective has to say.[30]

Consider the case of a change in price and quantity between the years 2014 and 2015. For the moment we're not going to necessarily assume that supply is perfectly elastic. In 2015, both the equilibrium price and quantity are higher than they were in 2014. Since demand curves must pass through both points, we know that demand must be higher in 2015 than it was in 2014. We also know that the change in demand (ΔD, measured in the quantity dimension as a percentage of the initial quantity) is greater than the change in quantity—this is evident from Figure 11-3.[31] The same graph for supply would imply that the change in supply (ΔS, also measured in the quantity dimension as a percentage of the initial quantity) is less than the change in quantity. In summary, if we let ΔP and ΔQ denote the percentage equilibrium changes in price and quantity, respectively, then, $\Delta D > \Delta Q > \Delta S$ when $\Delta P > 0$ and the opposite when $\Delta P < 0$.

[30] See Mulligan et al (2018, Section 2.2) for an example.

[31] Alternatively, the Δ operator can denote the change in the natural log. E.g., $\Delta Q \equiv \ln Q_{2015} - \ln Q_{2014}$.

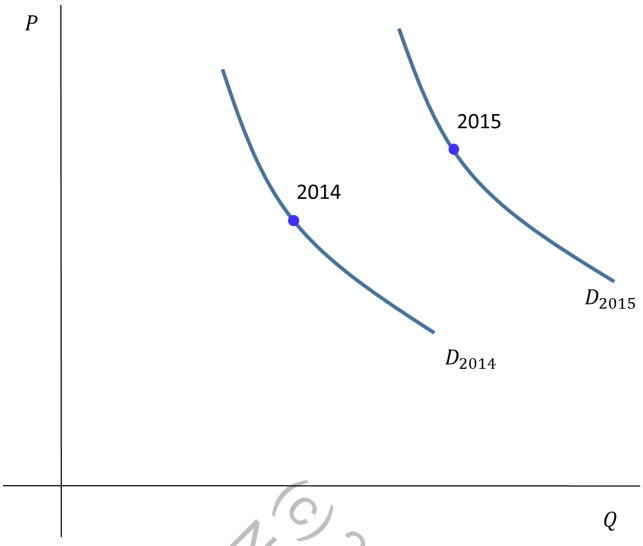


Figure 11-3: Since the demand curve intersects the equilibrium point, the demand curve must have shifted outward from 2014 to 2015. Moreover, the change in quantity is greater than the change in demand for each equilibrium price.

We can think about this example in terms of elasticities as well. Recall that $\Delta Q = \Delta D + \epsilon^D \Delta P$ which implies that $\Delta D = \Delta Q - \epsilon^D \Delta P$. That is, the change in quantity can be decomposed into a shift in the demand curve and a movement along the demand curve. Similarly, on the supply side, $\Delta Q = \Delta S + \epsilon^S \Delta P$, which implies that $\Delta S = \Delta Q - \epsilon^S \Delta P$. In both cases, ϵ denotes an elasticity of quantity with respect to price. In this example, we know ΔQ and ΔP, so, given also the elasticities of supply and demand, we could calculate exactly what the shift in demand and supply had to be.

Note that the equilibrium price and quantity are on both the supply curve and the demand curve. An equilibrium quantity must simultaneously be the quantity that sellers want to sell and the quantity that buyers want to buy. That is why we are keeping track of both a ΔD equation and a ΔS equation. This is a simple point, but people often forget one side of the market or the other.

Because ΔQ is shared in both equations, we can eliminate it and solve for ΔP as a function of the demand and supply shifts.

$$\Delta D + \epsilon^D \Delta P = \Delta S + \epsilon^S \Delta P$$

$$\Rightarrow \Delta P = \frac{\Delta D - \Delta S}{\epsilon^S - \epsilon^D}$$

This says that the percentage change in price is equivalent to the change in excess demand divided by the "sum" of the elasticities, since $\epsilon^D < 0$. Similarly,

$$\Delta Q = \frac{\epsilon^S \Delta D - \epsilon^D \Delta S}{\epsilon^S - \epsilon^D}$$

So the percentage change in quantity is a weighted average of the percentage change in demand and the percentage change in supply.

Four Ingredients of the Industry Model

In our constant-returns model of the industry, we can incorporate our additional equilibrium conditions $L = \frac{\partial C(w,r,Y)}{\partial w}$, $K = \frac{\partial C(w,r,Y)}{\partial r}$, and $Y = F(L,K)$. To summarize, the industry model therefore has 4 ingredients:

1. $P = MC$
2. $Y = D(P)$
3. $L = \frac{\partial C(w,r,Y)}{\partial w}$ and $K = \frac{\partial C(w,r,Y)}{\partial r}$
4. $Y = F(L,K)$

We can consider each of these in turn. The first equation corresponds, in the constant returns case, to $P = C(w,r,1)$. Totally differentiating, then $dP = C_w(w,r,1)dw + C_r(w,r,1)dr$. Because of constant returns to scale, then $dP = \frac{L}{Y}dw + \frac{K}{Y}dr$. Multiply and divide appropriately to convert changes to percentages: $\frac{dP}{P} = \frac{wL}{PY}\frac{dw}{w} + \frac{rK}{PY}\frac{dr}{r}$. This implies that $\Delta P = S_L \Delta w + S_K \Delta r$, where S denotes factor spending as a share of revenue and Δ denotes a percent change.[32] So the change in the price of output is a share-weighted average of the changes in input prices. As long as price equals marginal cost, this equation must be satisfied.

The second equation implies that $\Delta Y = \epsilon^D \Delta P$.

The third equation makes a new concept relevant: the elasticity of substitution, typically denoted σ. This measures how much factor inputs respond to changes in factor prices along the isoquant. Because of our previous discussion about the effects of a CRS assumption on isoquants (see Figure 11-1), we know that moving along a ray starting at the origin to different isoquants will maintain the same elasticity of substitution. More precisely, we will define $\sigma > 0$ to mean

$$\Delta \frac{L}{K} = \sigma \, \Delta \frac{r}{w}$$

There is a negative relationship between the ratio of factor quantities and the ratio of factor prices (recall that r is K's price, and w is L's). This condition implies that

$$\Delta L - \Delta K = \sigma(\Delta r - \Delta w)$$

The fourth equation can be treated in the same way we treated the first equation. $Y = F(L,K)$ implies that $dY = \frac{\partial F}{\partial L}dL + \frac{\partial F}{\partial K}dK$. This means that $\frac{dY}{Y} = \frac{P\frac{\partial F}{\partial L}}{PY}\frac{dL}{L} + \frac{P\frac{\partial F}{\partial K}}{PY}\frac{dK}{K}$, which yields $\Delta Y = S_L \Delta L + S_K \Delta K$.

Industry Elasticity of Labor Demand

The first confounding element one might consider when contemplating how a change in the wage affects labor demand is capital. For now, assume $\Delta r = 0$. This is typically a long run assumption. Then $\Delta P = S_L \Delta w$, $\Delta Y = \epsilon^D \Delta P$, $\Delta L - \Delta K = -\sigma \Delta w$, and $\Delta Y = S_L \Delta L + S_K \Delta K$. When the wage changes, we will

[32] Note that, by using the rules of calculus, we are considering small changes. Larger changes are examined by accumulating small changes, as shown in Chapter 4. The uppercase *S*'s refer to shares of revenue whereas lowercase *s*'s refer to shares of costs. They are the same with constant returns ($C = PY$); otherwise the *S*'s do not sum to one across factors, even though the *s*'s do.

see changes in price, output, labor, and capital, so we have four equations to solve for four unknowns. Using the first two, we can simplify to achieve

$$\Delta Y = S_L \epsilon^D \Delta w$$

This is the scale effect. An increase in the wage here drives output down. Now rewrite the fourth equation as $\Delta Y = \Delta L + S_K(\Delta K - \Delta L)$. This implies that $\Delta L = \Delta Y + S_K(\Delta L - \Delta K)$. This means the change in labor is going to be the scale effect *and* the substitution effect. We can use the third equation to plug in for $\Delta L - \Delta K$ to achieve $\Delta L = S_L \epsilon^D \Delta w + S_K(-\sigma \Delta w)$. The first term is the scale effect, and the second is the substitution effect. This can be rewritten to yield

$$\Delta L = (S_L \epsilon^D - S_K \sigma)\Delta w$$

The more elastic output demand, the more labor will decline. The more substitutable labor and capital are, the more labor will fall. These two are part of *Marshall's Law*. A third aspect of Marshall's law stated that increasing labor's share would drive labor further downward; the issue with this argument is that S_L and S_K are, of course, related, so increasing S_L also reduces the ability to substitute capital for labor (Hicks later corrected this law, noting that it additionally requires ϵ^D to have greater magnitude than σ).

Now consider, as an exercise, the short run elasticity of labor demand. That is, hold capital fixed so that $\Delta K = 0$. This will give the equations $\Delta P = S_L \Delta w + S_K \Delta r$, $\Delta Y = \epsilon^D \Delta P$, $\Delta L = \sigma(\Delta r - \Delta w)$, and $\Delta Y = S_L \Delta L$. Solving these equations in the same way as before will yield the short-run elasticity of demand for labor.

Are Labor and Capital Complements or Substitutes?

This can be restated: in the long run ($\Delta r = 0$), does $\Delta w > 0$ imply $\Delta K > 0$ or $\Delta K < 0$? In the short run ($\Delta K = 0$), does $\Delta w > 0$ imply $\Delta r > 0$ or $\Delta r < 0$?

Recall that the scale effect works in the following way: $\Delta w > 0 \Rightarrow \Delta P > 0 \Rightarrow \Delta Y < 0$, so the scale effect is always pushing in the direction of less labor and capital. The substitution effect means $\Delta w > 0 \Rightarrow \Delta \frac{K}{L} > 0 \Rightarrow \Delta K > 0$, so K is driven upward. So the question boils down to: is the elasticity of demand or the elasticity of substitution more important?

In the long run, using $\Delta L - \Delta K = -\sigma \Delta w$ and the formula for the change in labor, we have

$$\Delta K = (S_L \epsilon^D - S_K \sigma)\Delta w + \sigma \Delta w = S_L(\epsilon^D + \sigma)\Delta w$$

In the short run, $\Delta K = 0$, so the equilibrium equations are $\Delta P = S_L \Delta w + S_K \Delta r$, $\Delta Y = \epsilon^D \Delta P$, $\Delta L = \sigma(\Delta r - \Delta w)$, and $\Delta Y = S_L \Delta L$. These give that $S_L \sigma(\Delta r - \Delta w) = \epsilon^D(S_L \Delta w + S_K \Delta r)$, so $(S_L \sigma - S_K \epsilon^D)\Delta r = (\epsilon^D + \sigma) S_L \Delta w$.

$$\Delta r = \Delta w \frac{(\epsilon^D + \sigma) S_L}{(S_L \sigma - S_K \epsilon^D)}$$

Since $\epsilon^D < 0$, the denominator is positive. When $(\epsilon^D + \sigma) > 0$, the substitution effect dominates the scale effect so the price of capital rises in the short run and amount of capital rises in the long run.

Consider a subsidy to capital in the automobile industry. Will that lead to more or less labor in the industry? Assume a closed economy where the country produces and consumes all of its cars. The scale effect is likely to be smaller in this closed economy case because there are fewer substitutes in the output

market. Now suppose the state of Illinois decides to subsidize capital in the automobile industry. This would tend to make the scale effect bigger in Illinois, since the demand for Illinois-produced cars is more elastic than the demand for U.S.-produced cars (because we can move across markets). The effect of a capital subsidy on employment is thus different depending on the level at which it is supplied, state vs. national.

So the scale effect tends to make labor and capital complements, whereas the substitution effect tends to make them substitutes. Depending on the circumstance and horizon, the mix of substitution and scale effects can be different. This can make determining whether labor and capital are complements are substitutes a difficult question.

For another example, consider a pilot employed by a commercial airline. The pilot is a small share of the total cost for flights, so the demand for the pilot is fairly inelastic, in the sense that the scale effect will be small. But this means the complementary shares are large, so that an increase in the price of pilots will drive airlines to use bigger planes and more fuel. Marshall neglected the fact that an input with only a small share increases the ability for the firm or industry to substitute to other inputs, potentially increasing the magnitude of the substitution effect.

Considering More than 2 Inputs

The major concern in expanding to more than two inputs is how to think about the elasticity of substitution. We define a notion of a partial elasticity of substitution between factors i and j, called σ_{ij}. This is defined

$$\sigma_{ij} = \left(\frac{\partial^2 C}{\partial w_i \partial w_j} C\right) / \left(\frac{\partial C}{\partial w_i}\frac{\partial C}{\partial w_j}\right)$$

Where C refers to the cost function, and i and j refer to inputs i and j. In the three-factor case, for example, it now matters whether we increase the relative price for input i by increasing the price of input i or decreasing the price of input j.

Chapter 12 The Consequences of Prohibition

The Revenue from Drug Sales

To be concrete, let's consider a model of illegal drugs.[33] We want to think about there being a demand for drugs, and we want to initially assume they are legal and the industry is perfectly competitive and has constant marginal cost. Perfect competition is easy to relax, and the constant marginal cost seems to be a good assumption empirically. Now we will implement a prohibition on drugs and initially assume there is no effect on demand. The key point will be that the prohibition raises the cost of supplying drugs by forcing suppliers to do things they otherwise wouldn't. That is, they have to deviate from more efficient strategies. Raising the marginal cost raises the price. This reduces the quantity demanded. We'll assume demand is inelastic. For simplicity assume $\epsilon^D = -1/2$, and $\hat{p} = 4p_c$. Then $\hat{q} = \frac{1}{2}q_c$, since $q = Ap^{-\frac{1}{2}}$ is the constant-elasticity demand function. This means we are using twice as many resources as before to produce half as many drugs. Consumer expenditure is twice as much, going from E_c to $\hat{E} = 2E_c$ as in Figure 12-1. We must think that the externality from increasing consumption from $\hat{q}$ to q_c must be greater than the additional costs we incur *not* to produce that additional quantity.

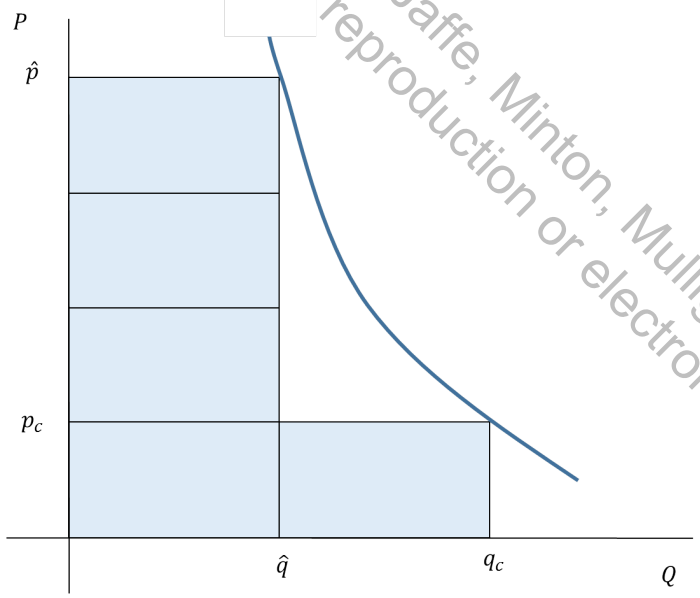


Figure 12-1: In this example, we use twice as many resources to produce half as many drugs.

Thus, this method of raising the real cost of consumption is inefficient unless it is believed that there is a huge negative externality from consumption of the additional drugs produced. But there are also other externalities imposed from making drugs illegal. For instance, drug dealers are often very violent, and the effects of this violence affect people who are not drug dealers. Further, drug dealers may corrupt local officials around the world, or fail to pay income, sales, excise, and other taxes, which impose negative externalities that are broader in scope.

The legal status of drugs can affect demand by making purchases less convenient or more socially stigmatized. We can allow for this, while still using Figure 12-1, by reinterpreting the price shown in the

[33] For further reading on this approach, see Becker and Murphy (2006).

figure as the full price. The full price is the sum of the money price paid to drug sellers and the additional inconvenience and stigma costs experienced by the consumer. We still have that prohibition delivers half the quantity at twice the cost, but now consumer expenditure on drugs measures only part of those costs. The willingness-to-pay-money schedule, not shown in the figure, is below the demand curve. Consumer expenditure and therefore industry revenue is measured as the equilibrium willingness to pay times the equilibrium quantity $\hat{q}$.

The legalization multiplier

Figure 12-1 shows a prohibition that fully eliminates legal activity E_c in the industry and replaces it with illegal activity that is twice as costly. Those extra resources come from elsewhere in the economy. Holding constant the supply of production factors to the total economy, and for simplicity assuming that the rest of the economy is legal, Figure 12-1 is showing that prohibition in this industry reduces legal activity outside of the industry just as much as it reduces in inside the industry, which itself is a lot. To the extent that legal activity is subject to income, sales, excise and other taxes, the reallocation from legal to illegal is a negative externality of the prohibition itself in addition to the violence, corruption, etc. already mentioned.

The supply of labor to the total economy may not be constant because the prohibition affects productivity (producing few drugs with more resources), although the effect of productivity on labor may be small and of ambiguous sign because productivity has both income and substitution effects. As we show in later chapters, less productivity probably means less aggregate capital.[34]

It follows that the legalization of drugs would not only expand the legal drug sector but also other legal sectors. Legalization therefore has a lot in common with productivity growth in agriculture or some other industry with inelastic demand: the industry produces more while freeing up resources for the rest of the economy to also produce more.

Half-hearted prohibitions are the most costly

Note that, while this analysis has been largely against the idea of a prohibition, prohibitions can be beneficial in some situations. If demand becomes elastic at low quantities (as it must for any demand curve with a choke point)[35] and enforcement is effective enough that quantity is almost fully reduced to zero, then even though the per-unit costs of the goods still produced will be exceptionally high, the total costs will still be low because they are incurred for so few goods. With this type of demand curve, prohibition is only a bad policy when it is not particularly effective. When it's not effective, prohibition imposes significant costs on suppliers and society at large. It is only worth it in these cases if there is a strongly negative effect of the additional goods being available.

In fact, we may simply want to exterminate an industry where we would not want merely a modest reduction. See Figure 12-2. Let's assume linear demand, which means that demand has a choke point near

[34] The consumption of drugs has a negative externality, which we have not specified in detail. Perhaps some of that negative externality is to reduce productivity in the legal sector.

[35] The choke point is the point where the demand curve intersects the vertical axis (that is, price has gotten so high that all demand is "choked off").

which demand is price elastic. Let's also assume a competitive industry constant marginal cost industry, which gives that total cost equals total revenue. Thus $P = D(Q) = C$ and $CQ = D(Q)Q$, where $D(Q)$ denotes the inverse demand curve. Government enforcement of a prohibition raises C and therefore P, and thereby indirectly determines consumption Q. The socially optimal consumption balances the production costs CQ, and any external costs, with the benefits to consumers.

Because, for any quantity, production costs are equal to industry revenue $D(Q)Q$, the marginal production cost of expanding quantity is equal to marginal industry revenue. Recall from the usual monopoly model that an industry's marginal revenue curve is below the demand curve, as shown in Figure 12-2.

Now suppose K is the per unit externality. We shift the demand curve (i.e., private marginal benefit curve) down by K to get the marginal social benefit. Note that for all the outputs between 0 and q^*, where the marginal cost line intersects the marginal social value line (D-K), marginal cost exceeds the marginal social value. In other words, at any quantity between 0 and q^*, society gains from further reductions in quantity. Conversely, at any quantity between q^* and the (unregulated) quantity q_c, society gains from further increases in quantity. Thus the quantity q^* where the two lines intersect minimizes social surplus.

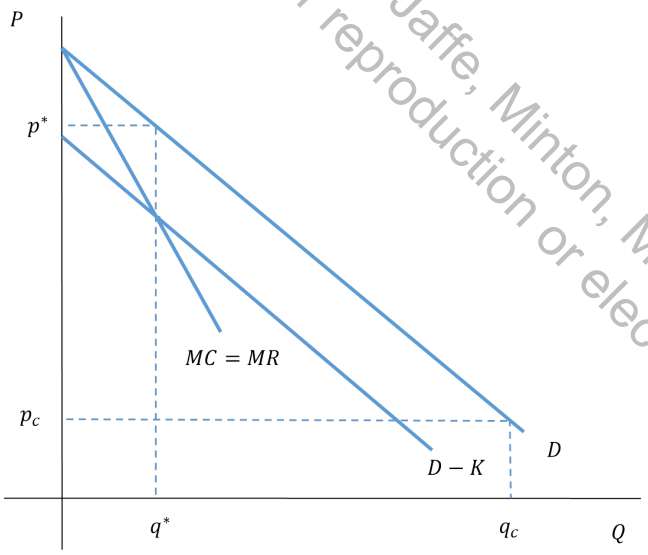


Figure 12-2: After accounting for the per unit externality K, it is optimal either to eradicate the industry or leave it at the competitive quantity and price. It is a minimum, in fact, at the pair (q^*, P^*) depicted.

So in this simple example, if we can set whatever price we want, we either want to do nothing and stay at the competitive output or completely eradicate the industry. Which one we choose to do depends on how large the externality is. We lose surplus on the first q^* units but then gain surplus as we move from q^* to q_c. We need to assess the net gain here. This process provides some intuition for why ineffective prohibitions can be very inefficient.

Chapter 13 A Price-theoretic Perspective on the Core

So far we have looked at cases where equilibrium price and marginal cost coincide. We noted in the context of firm theory that this case is more applicable than it first appears because marginal costs should be interpreted broadly. Also the equilibrium gap between marginal cost and price, if there is one, can sometimes be sufficiently constant that we get the correct comparative statics even when we take the gap to be zero. Because there are interesting behaviors for which a constant gap is not a satisfactory treatment, we use this chapter to show how a small adjustment to the previous analysis can add a lot of insight.

Looking for gains from trade: indifference curves for buyers and sellers

Recall Figure 5-1 where we drew a consumer's indifference curve and demand curve on the same quantity-price diagram, based on the fact that an individual's demand curve indicates the optimal quantity purchased at each price. Each point on the demand curve, all of the way up to the demand curve's choke point, has an indifference curve going through it.[36] The indifference curve going through the demand curve's choke point can be called the all-or-nothing demand curve. At any point on the all-or-nothing demand curve, the consumer is indifferent between bundles at prices along that curve and buying no units at the price where his demand curve intersects the y-axis. The consumer would rather not buy any units than purchase a bundle to the right of the all-or-nothing demand curve. Figure 13-1 adds that to what is already shown in Figure 5-1.

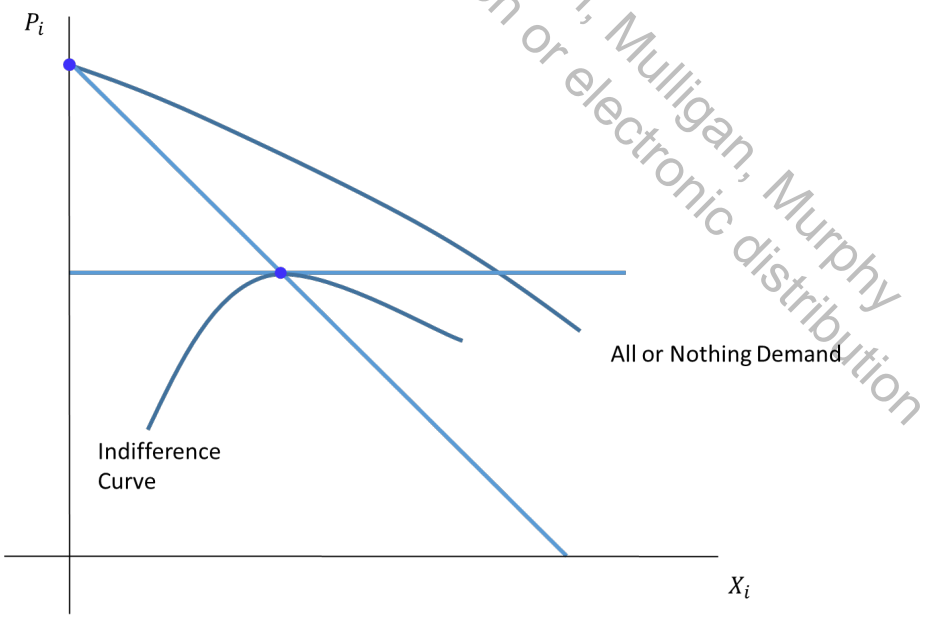


Figure 13-1: In (X_i, P_i) space, we can draw indifference curves. For any point along the all or nothing demand curve, for instance, the consumer is indifferent between purchasing a quantity at a price along that curve and purchasing no units at all at the price where his demand curve intersects the y-axis.

[36] The height of the choke point is the price at which demand is zero.

Now think about a firm that faces a downward sloping demand curve. He is a monopolist, so he calculates the marginal revenue curve. For simplicity, we assume that marginal cost is constant. The familiar result is that the monopolist constrained by the demand curve sets quantity where marginal revenue equals marginal cost, as at Q^* in Figure 13-2.[37]

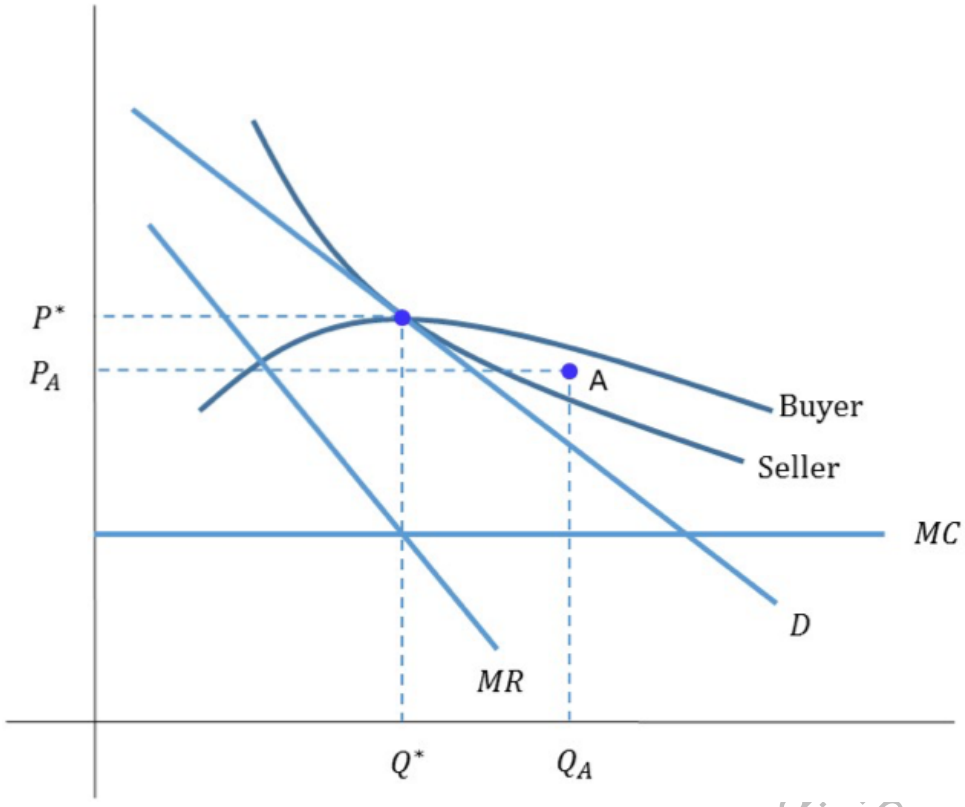


Figure 13-2: Both the monopolist and the consumer have an incentive to move from the monopoly solution to point A. Key features of point A are that the price is lower and quantity is higher than the market equilibrium, and the consumer is off the demand curve.

But we could get the same result by looking at the monopolist's "indifference curves" – more precisely his isoprofit curves, which are points where profits are equal. The isoprofit curves corresponding to positive profits slope down because getting the same profit at a lower price requires selling more units. Maximizing profits taking the demand curve as a constraint yields the profits associated with the isoprofit curve that is tangent to the demand curve, which is also shown in Figure 13-2.

Now draw the consumer's indifference curve at the market equilibrium, (Q^*, P^*), and note that there is a region where both the consumer and the monopolist can be made better off. Point A in Figure 13-2 has been drawn in this region.

[37] For the moment we assume that all buyers are identical so that we can show an individual's demand curve in the same picture with the monopolist's marginal revenue curve.

In this region, prices are lower than the equilibrium price, quantity is higher than the equilibrium quantity, and the consumer is off his demand curve. That is $P_A < P^*$, $Q_A > Q^*$, and $Q_A > Q^D(P_A)$. Note that both sides want to renege on this arrangement. After getting the lower price P_A, the consumer would rather purchase $Q^D(P_A)$. Similarly, after negotiating the higher quantity Q_A, the producer would rather receive revenues P^*Q_A. Thus, there needs to be some commitment here. One example of this scenario would be a negotiated discount, for instance.

This can be very good for consumers. Consumers are often willing to let grocery stores push them off their demand curves because this gives grocery stores more negotiating power with producers. In effect, they can make the consumer demand curve look more elastic by being able to push consumers off the demand curve in their stores.[38]

As explained by Klein and Murphy (2008), there is a similar reason for why so many fast-food chains serve Coke *or* Pepsi, rather than both. They have more negotiating power if they can tell Coke and Pepsi that they will lose all their business if they don't offer a deal. The ability to push people between Coke and Pepsi gives the fast-food chains this ability. And typically the price reduction achieved through this process compensates for the small losses consumers incur by being pushed around off their demand curves (i.e., getting more Coke than they'd like at the Coke-only restaurant and more Pepsi than they'd like at the Pepsi-only restaurants).

If buyers are heterogeneous, the monopolist constrained by demand may choose a price at which some of the buyers purchase nothing. Pepsi may choose a price that induces Coke lovers to drink no Pepsi. Here the all-or-nothing demand curve is relevant and shows how both the Coke lovers and Pepsi would be better off at a point below the all-or-nothing curve and above the Coke-lover demand curve. Such a point may be achieved when restaurants obtain a lower price for Pepsi by agreeing to serve Pepsi only. The Coke lover ends up drinking Pepsi when he dines at such restaurants, but pays less than he would at a restaurant that served both brands.

This perspective also helps us think about natural monopolies. Consider the structure and organization of the market as an outcome, not something that just happens. The problem with the natural monopoly model is it does not consider the cost curves in a broader context. Consider the case in Figure 13-3, where marginal cost is below average cost.

[38]The economic "theory of the core" makes a similar argument about gains from trade between buyers and sellers. See especially Telser (2006).

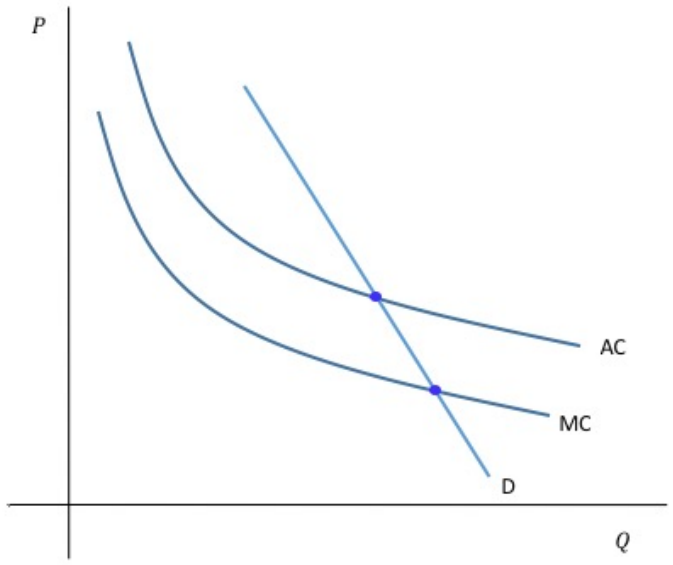


Figure 13-3: Marginal cost is below average cost, so the efficient outcome would mean firms are losing money on this good.

The natural monopoly model says that we cannot achieve the efficient outcome because firms would be pricing below average cost. But what if we bundle this product with something else? Perhaps it would be possible to sell the good at marginal cost if it is possible to bundle it with another good where the firm makes enough to offset the cost. This is what the supermarket does. The supermarket doesn't just sell ketchup, which would have average cost above marginal cost as a stand-alone product. Instead, the supermarket bundles many items together, realizing some of the gains from trade that, according to the natural monopoly model, would be lost due to the structure of ketchup costs.

Chapter 14 Multiple Factor Industry Model

Review of the Industry Model

Recall the two-input model of industry demand. We ended up with four equations:

$$S_L \Delta L + S_K \Delta K = \Delta Y$$

This is like a first-order approximation to the production function. In the Cobb-Douglas case, this *is* the production function. We also had

$$S_L \Delta w + S_K \Delta r = \Delta P$$

This was a first-order approximation to the cost function, assuming constant returns to scale. The first equation did not assume constant returns to scale—if we did not have CRS, we would just have that $S_L + S_K \neq 1$. This equation, on the other hand, will change in two significant ways. S_L and S_K will become marginal shares, and the equation will additionally become a function of Y, since changes in output will affect marginal cost when we do not have CRS. CRS really does two things: marginal factor shares become equal to average factor shares, because the industry uses inputs on the margin in the same way that it uses them on average, and the marginal cost becomes independent of output.

Note that homothetic production retains the feature that the isoquants increase radially out from the origin; they just are not necessarily proportional to the distance from the origin. CRS production gives a cost function along the lines of $YC(w,r,1)$, but homothetic production gives a cost function like $g(Y)C(w,r,1)$. Further, we had

$$\Delta Y = \epsilon^D \Delta P$$

This just says we will move along the industry demand curve. Finally, we had the substitution equation

$$\Delta L - \Delta K = \sigma(\Delta r - \Delta w)$$

In the two-factor case, or with a homothetic production function, the relative factor demands depends only on the relative price of those two inputs. There is also a generalization of the elasticity of substitution that is useful for the multiple factor case: the partial elasticity of substitution.

$$\sigma_{ij} = \left(\frac{\partial^2 C}{\partial w_i \partial w_j} C\right) / \left(\frac{\partial C}{\partial w_i}\frac{\partial C}{\partial w_j}\right)$$

Recall from our demand system analysis that second derivatives of the cost function (equivalently, price derivatives of the conditional demand functions) are restricted by adding up. In partial-elasticity-of-substitution format, adding up looks like:

$$0 = \sum_j s_j \sigma_{ij}$$

where the *s*'s are factor-spending shares.

Properties of the Multiple Factor Industry Model

Remember that

$$P = \frac{\partial C}{\partial Y}$$

$$Y = D(P)$$

$$X_i = \frac{\partial C(w_1, \ldots, w_N, Y)}{\partial w_i}$$

$$Y = F(X_1, \ldots, X_N)$$

How do we think about factor demand in this world? Consider, at the industry level and imposing constant returns to scale, the derivative $\frac{\partial X_i}{\partial w_j}$. Using the first equation, note that $\frac{\partial^2 C}{\partial Y \partial w_j} = \frac{dP}{dw_j} + \frac{\partial^2 C}{\partial Y^2}\frac{dY}{dw_j}$. The second term cancels, because under CRS, $\frac{\partial^2 C}{\partial Y^2} = 0$. Now we're left with $\frac{\partial^2 C}{\partial Y \partial w_j} = \frac{dP}{dw_j}$. But $\frac{\partial^2 C}{\partial Y \partial w_j} = \frac{\partial^2 C}{\partial w_j \partial Y}$, $\frac{\partial C}{\partial w_j} = X_j$, and $\frac{\partial X_j}{\partial Y} = \frac{X_j}{Y}$ by CRS. So $\frac{dP}{dw_j} = \frac{X_j}{Y}$. This means that the change in price due to a change in a factor price is just how much of the factor the industry uses per unit of output. This is related to the fact that CRS implies marginal shares are equal to average shares. Using the second equation, derive that $\frac{dY}{dw_j} = \frac{\partial D}{\partial P}\frac{dP}{dw_j}$. Finally, use the third equation to get that $\frac{dX_i}{dw_j} = \frac{\partial^2 C}{\partial w_i \partial w_j} + \frac{\partial^2 C}{\partial w_i \partial Y}\frac{dY}{dw_j}$. Finally, plug in for the relevant terms to achieve[39]

$$\frac{dX_i}{dw_j} = \frac{\partial^2 C}{\partial w_i \partial w_j} + \frac{X_i}{Y}\frac{\partial D}{\partial P}\frac{X_j}{Y}$$

As before, the left term on the right-hand side is the substitution effect, and the right term on the right-hand side is the scale effect. For $i \neq j$, in the two input case, the substitution effect is positive. In the multiple input case, if w_j increases, the substitution effect must on average be positive for other inputs. Now note that we can rewrite this entire equation in terms of elasticities:[40]

$$\frac{w_j}{X_i}\frac{dX_i}{dw_j} = \left(X_j w_j \frac{\partial^2 C}{\partial w_i \partial w_j} C\right) / \left(PY \frac{\partial C}{\partial w_i}\frac{\partial C}{\partial w_j}\right) + \frac{P}{Y}\frac{\partial D}{\partial P}\frac{X_j w_j}{PY}$$

$$\Leftrightarrow \epsilon_{ij} = s_j \sigma_{ij} + s_j \epsilon^D$$

As before, $s_j \sigma_{ij}$ is the substitution effect and $s_j \epsilon^D$ is the scale effect. Note that we can decompose the own-price elasticity in this way as well:

$$\epsilon_{ii} = s_i \sigma_{ii} + s_i \epsilon^D = -\sum_{j \neq i} s_j \sigma_{ij} + s_i \epsilon^D$$

So own-price elasticity is a share-weighted average of substitution elasticities and the output elasticity. The second equality is adding up for the second derivatives of the cost function.

[39] Another way to arrive that this condition is view it as part of a solution to a linear simultaneous-equations system. The system is formed by totally differentiating the (constant-returns versions of the) equilibrium conditions for p, Y, and X_i with respect to w_j, holding constant the other factor prices, and simultaneously solving for dp/dw_j, dY/dw_j, and dX_i/dw_j. Replace $\partial C/\partial w_i$ and $\partial C/\partial w_j$ with X_i and X_j, respectively.

[40] Here we have used $C = PY$ (constant returns) and $X_i = \partial C/\partial w_i$ and $X_j = \partial C/\partial w_j$.

This returns us to our discussion of Marshall's law from earlier in the book. There is no clear relationship between s_i and own-price elasticity. This is because some s_j change when s_i changes (s's sum to 1). Marshall noted that if a firm produces an intermediate input, it can raise the price significantly without much change in quantity if that intermediate input were a small factor for customers. This only incorporates the scale effect, however. If the substitution effect is more important, this is backwards, as Hicks pointed out. Recall that Marshall's other points were correct: if output demand ϵ^D is more elastic, then demand for the input is also more elastic. Further, own-elasticity of input i is negatively related to how substitutable i is with other inputs.

Analyzing Production

We can think about the production function $F(X_1, \dots, X_N)$ directly or the cost function $C(w_1, \dots, w_N, Y)$. These yield different first-order conditions. The first problem gives

$$P\frac{\partial F}{\partial X_i} = w_i$$

But the second problem gives

$$X_i = \frac{\partial C(w_1, \dots, w_N, Y)}{\partial w_i}$$

$$P = \frac{\partial C(w_1, \dots, w_N, Y)}{\partial Y}$$

For different problems, it will be useful to use different methods. If we want to hold quantities of other inputs constant, it is convenient to use the production function directly. If we want to hold prices of other inputs constant, it is convenient to use the cost function. In practice, what we hold constant corresponds to different experiments.

Endogenous Factor Prices

What we just did assumes that when we change w_i, all other prices remain constant. Consider the following model. We have a supply function of X_j, $X_j^S(w_j)$, which is upward sloping. Now we have the equilibrium conditions

$$X_j^S(w_j) = X_j(w_1, \dots, w_N, Y)$$

$$X_i = X_i(w_1, \dots, w_N, Y)$$

$$P = \frac{\partial C(w_1, \dots, w_N, Y)}{\partial Y}$$

$$Y = D(P)$$

Now, w_j is endogenous. As we change the price of factor i, we now allow w_j to change. For a purely short run analysis, make the elasticity of supply 0, so that $X_j^S(w_j)$ is fixed.

Just like in the consumer problem, there are multiple ways to think about the world. In the consumer problem, we had the Marshallian approach and the Hicksian approach. One is easy for some questions, and the other is easy for different questions. We can use the same kind of logic here. For this problem, it would be easier to use the production function approach, because taking partial derivatives of the cost function holds prices constant.