Module 4: Costs, Cost Curves, and Perfect Competition

The following is an excerpt from Ariel Rubinstein (2012), *Economic Fables*. It is reproduced without alterations. Full text available at https://www.openbookpublishers.com/product/136.

The tale of the three tailors

Imagine an island with six hundred residents, all dressed in identical clothes that require mending every month. Three tailors work at mending the clothes. For as long as anyone can remember, the residents of the island have been divided equally between the three tailors. Once a month, each resident goes to the same tailor whose services his father had used. Tradition, or decree, has set the price of the monthly repair at $5. Assume that the tailors have only minimal, negligible expenses. Each of the tailors would like to have as many customers as possible. However, even with great effort, none of them can do more than three hundred repairs a month. The residents feel that there is "hidden unemployment" in the tailoring sector. The tailors are often seen reading a newspaper, or dozing. It seems that two tailors would be enough and that it would be better if one of the tailors were to quit tailoring and find himself another job. In the language of economists, the situation is inefficient.

Let us say that all the tailors have various other employment options that influence their decision about whether to remain in the tailoring profession or to quit. Tailor A can expect to earn $900 a month in another profession, while Tailor B can expect to earn $600. Tailor C has limited alternative employment options and can earn only $300 outside the tailoring field. Each of the tailors will choose to abandon his sewing needles if his income from tailoring falls below his alternative income ("opportunity cost"). Currently, when the price of mending a piece of clothing is $5, it is not worthwhile for any of the tailors to leave this line of business because each tailor has two hundred customers and a monthly income of $1,000.

One day, the idea of the free market reaches the island. The traditions are shattered and the decrees canceled, and each tailor can decide on the price he charges for repairs. Each resident compares prices and turns to the tailor who offers his services at the lowest price. If more than one tailor offers the lowest price, the residents of the island will divide their custom equally between them. Each tailor attends a short course in modern business management and internalizes his role in the new economic regime: he must become familiar with the market and maximize his profits. What will happen on the island in the new situation?

The continuation of the Tale of the Three Tailors must provide answers to the following questions: Which tailors will remain in this occupation? What will be the terms of commerce between the tailors and their customers? As is customary in economics in this type of context, we will use a solution concept called **competitive equilibrium**. The concept of equilibrium imposes the following requirements for the rest of the story: (1) All customers will pay the same price for the repair of his clothes. (2) Each tailor knows the price of the service and compares the income he believes he can make in this work and his potential income outside this field. If the income in the other profession is higher, the tailor will leave the tailoring business. If the income outside of this field is lower, he will remain a tailor. (3) The number of customers the remaining tailors are interested in serving (supply) is equal to the number of islanders interested in this service (demand). Now, all six hundred islanders are interested in the service at any price. Since the tailors have no expenses, each one is interested in serving three hundred customers (the greatest number of repairs he is capable of doing a month). Thus, this condition demands that precisely two tailors remain in this business.

The logic underlying the concept of competitive equilibrium is that if the price of tailoring services is so high that the supply of tailoring services exceeds the demand, then the price will decline until one of the tailors closes his business. And if the price is low and the demand for tailoring services is greater than what the tailors are able to supply, the price will rise until another tailor returns to this sector.

We will now see that there is competitive equilibrium when the price of a repair is $2.50 (or any other price between $2 and $3), and only tailors B and C remain in this business sector. Each of the tailors (B and C) will have three hundred customers and each has an income of $750, which is more than either could receive in his alternative employment. Tailor A, meanwhile, earns $900 outside of the tailoring business. If he returns to this sector, he would earn $750 at most, less than what he is earning in another occupation.

In every competitive equilibrium, the price of the tailoring service will be lower than the price that prevailed in the old regime: If the price of the service were $5 (or higher), the tailor who quit the profession would figure that he could earn more as a tailor than he does in his new line of work. Only the two tailors whose alternative employment options are less profitable will remain in the tailoring sector; and the total output of the residents of the island will grow. An “invisible hand” generates the competitive equilibrium price and mobilizes the self-interest of the tailors and the islanders to correct the inefficiency created by the traditions and decrees that were recently canceled.

How does the market arrive at the competitive equilibrium price? The usual explanation offered in Introduction to Economics classes goes like this: The price of mending clothing prior to the cancellation of traditions and decrees was $5. After canceling the traditions and decrees, a price war erupts. One of the tailors who was “re-educated” concludes that it would be better for him to lower the price to $4.90 and thus create a situation in which all of the islanders would seek his services. Before long, the other tailors take note and also lower prices. And thus the price drops lower and lower until a certain stage when one of the tailors offers the service at a price less than $4.50. At this point, the tailor with the best employment alternative closes his tailoring business and engages in a different profession, and the island remains with only two active tailors.

Several assumptions in this story are not obvious. First, is it indeed so clear that the tailors will lower their prices after the cancellation of the traditions and decrees? We expect them to act only in pursuit of their own personal interests. But if a tailor is concerned only with his own earnings, it would actually be better for him not to lower the prices because he understands that any profit he would gain from increasing his clientele would be temporary and insignificant compared to the large loss he would incur in the future when the other tailors respond to this move and also lower their prices. The tailor would not need to speak with his colleagues in order to refrain from lowering prices. (Explicit collaboration between the tailors might be prohibited on the island under antitrust legislation.) Stated simply, no tailor would want to start a price war.

Second, let's assume that the tailors are not so wise and fall into the trap the competitive atmosphere lays for them. Is it clear that the consumers will indeed choose the least expensive tailor? Until now, they have used the services of the same tailor their father and grandfather used. Now they need to compare prices frequently. If the price differentials between the tailors are not large, some customers will decide that the price savings are not worth the bother involved in comparing prices. Thus, a tailor may actually raise his price a little, relying on the fact that most of the customers will not bother to find another tailor offering the service at a lower price. If customers do not compare prices, the market might stabilize at a higher price than the competitive equilibrium price.

Finally, let us assume that all of the residents of the island regard the search for the least expensive tailor as a real national mission, an act that will serve the society as economists demand, and let us assume that the tailors are not so smart, and that price competition rages and leads to a drastic drop in prices, and that one of the tailors abandons the profession and finds alternative employment (and does not become jobless on the streets of the island), and that it enlarges the national pie. Then we come to the question: is this story as happy as it sounds?

The change generated by the competitive economic regime on the island did indeed lead to growth in the "national pie." However, the improvement also led to a change in the distribution of income. The situation is worse for the tailors and better for their customers. Is the income distribution better now? Are the tailors now receiving fairer compensation for their work? Is the price for mending clothes now more reasonable? There are no objective answers to these questions. Economics has no way of choosing between the new situation and the previous situation. The island's residents, all of them, are the ones who must make the choice.

Modeling Firm Costs

In the supply and demand model of a single market, we said the supply curve traced out a "marginal willingness-to-accept" curve for those who have the good and are interested in selling it. In some instances, these willingness-to-accepts will be driven by the preferences of the current owners of the goods. For instance, a homeowner might have a willingness-to-accept for their home that is higher than the assessed value of the home because they have lived in it for some time and they are comfortable with it. Or a potential worker thinks about the alternative uses of their time when they decide whether to work and if so, how much.

That said, many goods are produced by firms (companies or businesses) that purchase inputs on markets and convert those inputs into something which they sell to consumers. These firms consider their costs (both monetary and non-monetary) when deciding whether to produce, and if so, how much to produce. Their ultimate goal is to make a profit, that is, take in more revenue than it costs to produce. Many firms operate in an environment where most of their costs are paid out in dollars. This makes it easier to see whether they are turning a profit, since the accountants can answer that question using a spreadsheet. As we will see, the economists' cost concepts implore us to also pay attention to costs that are not on the books, as sometimes they are substantial.

Because a firm's marginal willingness-to-accept, and thus its supply curve, depends on its cost structure, we are motivated to think more about firm production cost curves, that is, how firm costs are related to output. Even apart from this motivation, the cost concepts we will learn also carry over to other organizations, including nonprofits and governments, which have other goals besides profit-maximization, but nevertheless need to consider costs in their decision-making.

The Economic Concepts of Cost

We'll start our discussion with some important cost concepts: fixed versus variable costs, opportunity costs versus accounting costs, private versus social cost, and sunk versus recoverable costs.

Fixed versus Variable Costs

A **fixed cost** is a cost that does not change as output changes. For example a firm might need to pay for the lights to be on in order for the workers to see what they are doing and for production to happen. But the lights are simply on or off and the cost of powering them does not change when output changes. Or a transportation bureau might need to build a bridge in order to provide river-crossing services, but once the bridge is built the construction cost does not vary with the number of cars that cross over it.

A **variable cost** is a cost that changes as output changes. For example a firm that wishes to produce more output might need to employ more labor hours by either hiring more workers or have existing workers work more hours. The cost of this labor is therefore a variable cost as it changes as the output level changes. A transportation bureau, once it has built the bridge, might need to hire more toll collectors if more cars cross, and it will likely have to re-pave the bridge more often, which we can consider as a variable cost.

Opportunity Costs versus Accounting Cost

The **opportunity cost** of something is the value of the next-best alternative given up in order to get it. This includes explicit costs, paid for with money, and implicit costs, which are not paid in money. This contrasts with the **accounting cost** which considers *only* costs that are paid for with money, or explicit costs.

A classic example that distinguishes these two is an ostensibly free input for a firm or organization. Suppose a firm has access to an input that it can use in production without paying a price for it. A simple example is a family farm. The farm uses land, water, seeds, fertilizer, labor, and farm machinery to produce a crop—let's say corn–which it then sells in the marketplace. If the farm owns the land it uses to produce the corn, do we then say that the land component is not part of the firm's costs? If you used accounting cost, the answer would be, yes, because it wouldn't show up on the books. But the economist's answer is no, absolutely not. When the farm uses the land to produce corn it forgoes any other use of the land; that is, it gives up the *opportunity* to use the resource for another purpose. In many cases the opportunity cost is the market value of the input, but, in the case of ostensibly free inputs, it is not actually paid in dollars. A proper evaluation of costs for decision-making must consider opportunity costs.

For example, suppose an alternative use for the land is to rent it to another farmer. The forgone rent from the decision to use the land to produce its own corn is the farm's opportunity cost and should be factored into the production decision. If the opportunity cost, which in this case is the rental fee, is higher than the profit the farm will earn from producing corn, the most efficient economic decision is to rent out the land instead, and stop growing corn!

Opportunity costs are always higher than accounting costs, because they include implicit costs in addition to explicit ones. When we talk about profits in this course, we mean revenues minus opportunity costs, not minus accounting costs. We sometimes will say "economic profits" in order to emphasize that we're thinking of profits after deducting opportunity costs, not just accounting costs. This will be confusing because economic profits are usually not written down anywhere; instead, the accountants will be recording only accounting profits. But often economic thinking can help out by reminding folks that there alternative uses for some things that are not expressed in the accounting books. Rational actors should consider the full opportunity costs when deciding on a course of action, not just the accounting costs.

Now consider the more complex example of a farm manager who is told to produce a certain amount of corn. Suppose that the manager figures out that she can produce exactly that amount using a low-fertilizer variety of corn and all of the available land. She also knows that another way to produce the same amount of corn is with a higher yielding variety that requires a lot more fertilizer but uses only 75% of the land. The additional fertilizer for this higher yielding corn will cost an extra $50,000. Which option should the farm manager choose?

Without considering the opportunity cost she would use the low-fertilizer variety of corn and all of the land, because it costs $50,000 less than the alternative method. But what if, under the alternative method, she could rent out for $60,000 the 25% of the land that would not be planted? In that case the cost minimizing decision is actually to use the higher yielding corn variety and rent out the unused land.

Another classic example is that of a small business owner who runs, say, a coffee shop. The inputs into the coffee shop are the labor, the coffee, the electricity, the machines and so on. But suppose the owner also works a lot in the shop. He does not pay himself a salary but simply pays himself from the shop's excess revenues, or revenues in excess of the cost of the other inputs. The opportunity cost of his labor for the shop is not $0 but the amount he could earn working elsewhere instead. If, for example, he could work in the local bank for $4,000 a month then the opportunity cost of his working at the coffee shop is $4,000 and if the excess revenues are less than $4,000 he should close the shop and work at the bank instead, assuming he likes both jobs equally well.

One example often cited in nonprofit management is when universities own property in the center an expensive city (think NYC or LA), or other expensive locales. The organization might consider that property a "free" input into its production of educational output. Actually, if the organization could sell or rent out the property for a large sum of money, using it to give lectures might constitute a very expensive input. In some cases, if the goal of a university is to produce a large quantity of education at low cost, it might be optimal for the university to sell off or rent its historical locations and move away from the city. But if the university administration wrongly writes off its urban properties as "free resources," then these options may never be seriously considered. Of course, universities likely have different goals than simply minimizing cost per graduated student, and holding on to urban properties might boost the prestige of the university.

Notice how the idea of opportunity cost neatly captures what it means for an action to be "costly." An action is costly when by choosing it, the actor foregoes a lot of alternative things that are desired. If the university uses the expensive urban property for lectures, it is giving up a lot of money it could get from selling the property, moving to the countryside, building new buildings, hiring new faculty, and producing far more lectures and research. An economist would say that using the building for lectures is very costly – but nothing would show up on the books!

Private versus Social Costs

Assuming that I've convinced you that opportunity costs are the ones we should be concerned about when modeling the behavior of rational actors (and evaluating our own organization's decisions), let's consider two different types of opportunity costs: private and social (opportunity) costs.

The **private opportunity cost** of an action is the opportunity cost from the perspective of the actor who has the ability to take the action, or in other words, it is the reward (e.g. payment) to this actor needed to make the actor take the action. This includes the price of resources that the actor must purchase on the market to take the action as well as the implicit costs not part of accounting costs (that together form the opportunity cost of the action). The **social opportunity cost** of an action is the opportunity cost to all people in a society (or the world) from the actor taking the action. It is what, hypothetically, we would have to compensate everyone in society to permit the action to be taken. Since there is no specific individual we can ask who can tell us what the social opportunity cost of an action is, we can and do argue over what social opportunity costs entail. But the idea of there *being* "social opportunity cost" separate from "private opportunity cost," and sometimes these not being equal, is well-accepted in economics.

The private opportunity cost of an action may, and arguably often does, roughly equal the social opportunity cost of the action. If prices on the market accurately reflect the opportunity cost of using those resources for the action (including the implicit costs not on the books, such as the foregone rent from the university building discussed earlier), then when the actor makes the decision to do something, the costs they consider are, in fact, the social costs. The most common story we have where social costs deviate from private costs is when the prices do not incorporate the cost of the decision on people not involved in the decision. Environmental pollution is the classic example. When a company decides to produce a chemical that has the side-effect of polluting the neighborhood, and the company does not have to pay for the negative effects of the pollution, then the company's private cost of taking the action is less than the social cost of the action. Intuitively, the company will take this action too often from "society's perspective."

Later in the course we will spend much time talking about the distinction between private and social cost, since this split leads many to call for government action to rectify the resulting inefficiencies – if the market is left alone, too much pollution will occur, for example, so government should stop that from happening. For now, our analysis will assume private and social costs are equal.

Sunk versus Recoverable Costs

Some costs are **recoverable** and some are not. An example of a recoverable cost is the money a farmer spends on a new tractor knowing that she can turn around and re-sell it for the same amount she paid.
A **sunk cost** is an expenditure that is not recoverable. An example of a sunk cost is the cost of the paint a business owner uses to paint the leased storefront of his coffee shop with his shop's logo. Once the paint is on the wall it has no resale value (nobody cares about the logo *per se*). Many inputs reflect both recoverable and sunk costs. A business that buys a car for the use of an employee for $30,000 and can resell it for $20,000 should consider $10,000 of its expenditure a sunk cost and $20,000 a recoverable cost.

Why does the sunk versus recoverable cost distinction matter? Because after incurring sunk costs, their value as costs no longer matters when making future decisions. To be clear,

1. The size of to-be-sunk costs matters *before* making the decision that entails a sunk cost, because those costs are not yet sunk.
2. The consequences of making the sunk investment decision can and will affect future decisions.
3. But the amount of the sunk cost does not affect future decisions *per se*.

When a person or organization does not behave according to point (3) in this list, we say that an error has been made due to the **sunk cost fallacy**.

To give an example, suppose you buy a $500 non-refundable and non-transferrable airline ticket to go to Florida at spring break. However, as spring break approaches you are invited by friends to spend the break at a mountain cabin in Colorado they have rented to which they will give you a ride in their car at no cost to you, and you don't have to pay any of the cabin rent. You prefer to spend the break with your friends in the cabin, but you have already spent the $500 on the ticket and you feel compelled to get your money's worth by using it to go to Florida. The notion of *sunk cost* compels you to ignore the number $500 when making your decision. Not ignoring it may lead to a sunk cost fallacy.

To make this precise, first suppose we consider your decision *before* you purchased the nonrecoverable Florida ticket – let's imagine your friends told you about Colorado sooner for instance – your decision would be between these two options:

1. Happiness from Florida (WTP) - $500 Ticket
2. Happiness from Colorado (WTP)

Clearly the cost of the ticket matters in this decision. If it's $500, as written, that's one thing, but if it's $2000, that tips you more toward Colorado.

Now consider your payoffs at the current point of deciding between going to Florida or Colorado. You *already* paid for the nonrecoverable ticket to Florida. So, if we sum up your total payoffs from when you purchased the nonrecoverable Florida ticket to now, we get the following payoffs depending on your choice

1. Happiness from Florida (WTP) - $500 Ticket
2. Happiness from Colorado (WTP) - $500 Ticket

This comparison makes it clear that the best choice is now found by just comparing your WTP for Florida or Colorado. That sunk cost number, $500, is not relevant to your decision.[1] If it were $2000, or $200, it would

[1] I thought of a way to contradict this statement when writing this. Specifically, imagine a toy model in which you don't know or you forgot your WTP for Florida, but because in the past you decided to purchase the Florida ticket (after a great amount of thought about it perhaps) without knowing the other option, that tells you that your WTP for Florida must be larger than $500. Then when you try to remember what your WTP for Florida is at this point of decision between Florida and Colorado, your prior belief about your WTP is distributed according to $\Pr WTP$, but then you incorporate the fact that the ticket was $500, which gives you a conditional belief about your WTP of $\Pr WTP \geq \$500$,

not make a difference, your decision is now over Florida for free versus Colorado for free.

The sunk cost fallacy occurs when, at the current point of deciding, you calculate your payoffs like this

1. Happiness from Florida (WTP)
2. Happiness from Colorado (WTP) - $500 Ticket

or like this

1. Happiness from Florida (WTP) + $500 Ticket
2. Happiness from Colorado (WTP)

or in some other way such that your calculations don't capture that the $500 unrecoverable ticket cost is the same regardless of what you choose now. The first one is you saying, "if I go to Colorado then I lose the $500 from the ticket." The second is you saying, "if I go to Florida, I get my $500 worth." Either way, these calculations are wrong.

Often it is the "get my money's worth" phrase that suggests to me the sunk cost fallacy. Sure, the fact that you bought the ticket means the current decision is Florida for free versus Colorado for free – that is point 2 in the list above – of course decisions made in the past affect current decisions, because now you have a free trip to Florida on the table. But you don't "get $500 back" by going to Florida or "lose $500" by going to Colorado. Of course, you might feel regret for your past decision if it turned out

because your past decision tells you your WTP is above $500. I can see intuitively how this could lead the current decision to depend on the ticket price. This would be a "bounded rationality" model: the consumer is assumed rational but without full knowledge of their own preferences. I haven't precisely written down this toy model and I wouldn't make any guaranteed statements about its predictions until I saw it written down precisely. In any case, we don't really have the tools to think this way at this point in the course, because this requires uncertainty and belief updating. And this is a type of uncertainty that is unusual in economics, too – it assumes the consumer is uncertain about their own preferences. Economists do assume these sorts of thing in the models used in behavioral economics. We'll play with some models like this when we get to the uncertainty module (probably not this one specifically, though). For now, we are thinking the consumer knows their own WTP for Florida and Colorado with 100% certainty. If that's the case, it is clear that the consumer should discard the $500 ticket price information when making this decision, and if they don't, they might make a mistake. That's the whole sunk cost fallacy idea.

suboptimal, but really, people should not regret past decisions made after careful consideration with the information available at the time. The fact that they nevertheless do, and why, is a topic that I believe psychologists are better equipped to discuss, though I'm sure there are some models that behavioral economists have used to think through regret, too.

Cost Curves

Short-Run Cost Curves

Cost curves represent the relationship between a firm's output and the different cost measures involved in producing the output. Cost curves are visual descriptions of the various costs of production. In order to maximize profits, firms need to know how costs vary with output, so cost curves are vital to the profit maximization decisions of firms. They are also important for modeling the behavior of nonprofits and governments which, although they do not maximize profits, must balance revenues (fees, donations, and taxes) with cost in order to stay operational.

The structure of the costs of production has many implications for firm behavior, but ultimately these costs will determine each firm's willingness-to-accept, and thus the ultimate supply curve for a market. That being said, often in microeconomic storytelling, we will assume some rather simplistic form for cost, such as "constant marginal cost."

When we draw a cost curve that shows a single cost for each output amount, we are implicitly assuming firms optimally make decisions about which bundle of possible inputs to use so that costs are minimized. In this course, we abstract away from the input choice decisions of firms (and other organizations) as they select the optimal quantities of land, materials, labor, capital (e.g. machinery, buildings), etc., in order to produce output at a minimum cost. We do this in the interest of brevity. In the discussion to follow, we will occasionally refer to *inputs to production*. Hopefully, these can be understood intuitively: for a firm to produce output, it must pay for inputs, such as workers, machines, buildings, etc. We will sometimes refer to two specific categories of inputs talked about often in economics: labor (human workers) and capital (machines, buildings, etc.). We will assume that the firm can sometimes ("long-run") adjust all the inputs, and sometimes ("short-run") the firm cannot adjust some of the inputs. In this section, we are considering the short-run. Naturally, the firm's relative ability to adjust will affect the relevant costs of production, as we will discuss.

I note at this point for your knowledge only, that the theory of cost-minimization for a firm facing constant market prices (linear pricing) for inputs looks analogous to consumer theory, except with the moving pieces reversed: in consumer theory, a consumer faces a budget constraint and attempts to achieve the highest indifference curve given that budget constraint; in cost-minimization theory, the firm faces a production constraint and tries to achieve the minimum cost curve (which looks like a budget constraint due to linear pricing) given that production constraint. Cost-minimization can thus be seen as a tangency condition, with ramifications similar to consumer theory. The fact that cost-minimization and utility-maximization look like mirror images is referred to as "duality" in economic theory. However, these aspects of producer theory will not be further covered or tested in this course.[2]

We will consider two categories of cost curves: short-run and long-run. In this section we focus on short-run cost curves.

The Seven Short-Run Cost Measures

The short-run is a time period short enough that some inputs to production (e.g., machinery) are fixed while others (e.g., labor) are variable. There are seven cost curves in the short-run: fixed cost, variable cost, total cost, average fixed cost, average variable cost, average total cost, and marginal cost. Don't worry, they are related to each other in very intuitive ways. Moreover, it is the existence of "fixed costs" that is implied by the assumption that it is the "short-run:" the firm cannot adjust its fixed costs in the short-run.

[2] Modeling production plays an important role in applied public policy analysis. Analysts are very often concerned with the very issue of minimizing cost/maximizing the impact of dollars spent, which often requires considering government/nonprofit production functions. Moreover, macroeconomic models often use aggregate production functions. That we skip this in this course is not to downplay its importance. It is simply convenient that the theory of perfect competition, and theories later in the course, can be studied by taking cost curves as the primitives, rather than production functions. This helps us fit our course into the available time constraints.

The **fixed cost** (**FC**) of production is the cost of production that *does not vary with output level*. The fixed cost is the cost of the fixed inputs in production, such as the cost of a machine (capital) that costs the same to operate no matter how much production is happening. An example of such a machine is a conveyor belt in a factory that moves a car chassis through various stages of an assembly line. The belt is either on or off, but the cost of running it does not change depending on how many cars it carries at any point in time.

Note that the fixed cost of a piece of capital equipment that the company owns includes the opportunity cost as well. In our example, the fixed cost includes the actual costs of running the conveyer belt, such as power and maintenance, as well as the opportunity cost of using it for the firm's own production rather than renting it out to another company.

The **variable cost** (**VC**) of production is the cost of production that *varies with output level*. This is the cost of the variable inputs in production, for example the cost of the workers that assemble the electronic devices along a conveyor belt. The number of workers might depend on how many devices the factory is trying to produce in a day. If its production target increases, it uses more labor. Thus the hourly wages it pays for these workers are a variable cost. Variable cost generally increases with the amount of output produced.

Like fixed cost of production, there is an opportunity cost associated with variable cost of production. In this case, the next best alternative use of these workers is to go to another firm that will employ them. Thus the market wage for workers represents their opportunity cost, and as such, the wage cost of employing them is equivalent to their opportunity cost.

The **total cost** (**TC**) of production is the sum of fixed and variable costs of production:

$TC = FC + VC$.

This equation decomposes total cost into two components. Corresponding to these three costs, there are three short-run average cost measures: the average variable cost, average fixed cost and average total cost. These are just the previous cost measures divided by output.

Average variable cost (**AVC**) is the variable cost per unit of output. Mathematically, it is simply the variable cost divided by the output:

$AVC = VC/Q$

Note that since variable cost generally increases with the amount of output produced, the average variable cost can increase or decrease as output increases.

Average fixed cost (**AFC**) is the fixed cost per unit of output. Mathematically, it is simply the fixed cost divided by the output:

$AFC = FC/Q$

Because fixed cost does not change with the amount of output produced, the average fixed cost *always decreases* as output increases. In fact, AFC has a very specific functional form that I encourage you to plot (try plotting e.g. 1000/Q for Q from 0 to 100).

Average total cost (**AC**), or simply **average cost**, of production is total cost per unit of output. This is the same as the sum of the average fixed cost and the average variable cost:

$AC = AC/Q = AFC + AVC$

Because it is the sum of the average fixed cost, which is always declining with output, and the average variable cost, which may increase or decrease with output, the average total cost may increase or decrease with output. We typically draw AC as if it first decreases in quantity, as the AFC is spread over units, and then eventually increases as variable costs increase at an increase rate. This is not necessarily the case, though.

Marginal cost (**MC**) is the additional cost incurred from the production of one more unit of output. Marginal cost is very important for understanding the behavior of the firm, as it also the firm's marginal willingness-to-accept (perhaps more accurately we could call it the firm's marginal willingness-to-produce). In notation, marginal cost is

$MC = \Delta C/\Delta Q \; for \; \Delta Q = 1$

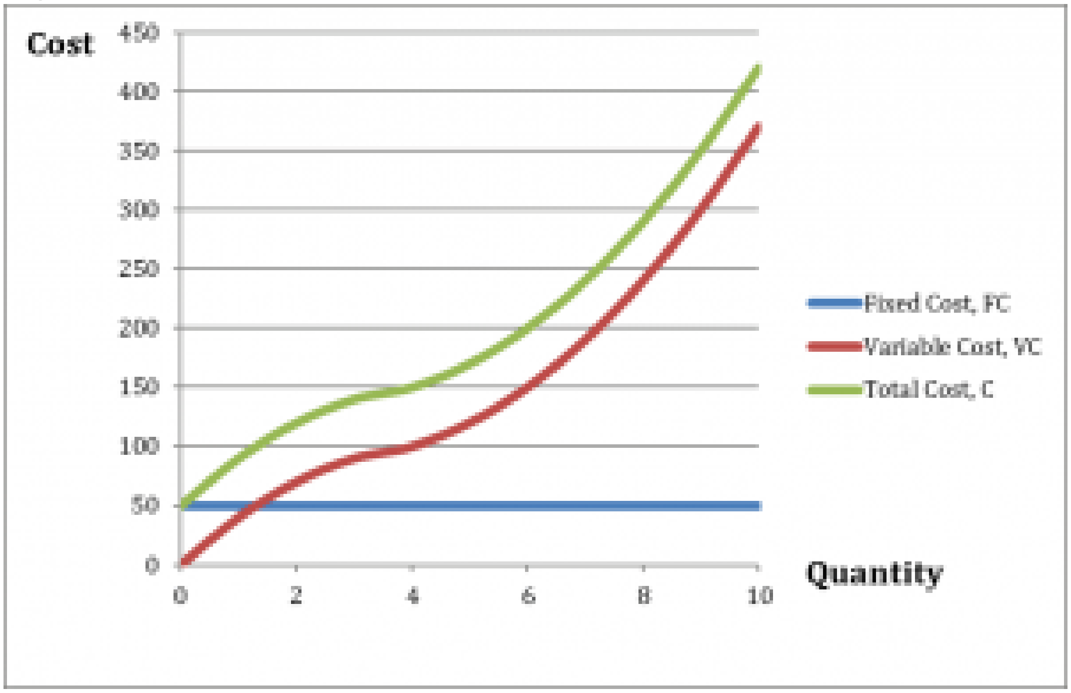
Since the only part of total cost that increases in quantity produced is variable cost, we can equivalently express marginal cost as

MC = ΔVC/ΔQ $for\ \Delta Q = 1$

Fixed Cost, Variable Cost, and Total Cost Curves

To give a concrete example, the Figure below summarizes a firm's daily short-run costs. Perhaps this firm is a small air conditioning repair business and the quantity is the number of repairs done per day. The firm has a fixed cost of $50 per day of production. This cost is incurred whether the firm produces or not, as we can see by the fact that $50 is still a cost when output is zero. The firm's fixed cost of $50 does not change with the amount of output produced. Perhaps this fixed cost includes the opportunity cost of the tools and machinery owned by the firm, and the cost of the truck (including cost of maintenance plus the opportunity cost of the foregone interest that could be earned from selling the truck).

Figure: Three short-run cost curves – fixed, variable and total costs.

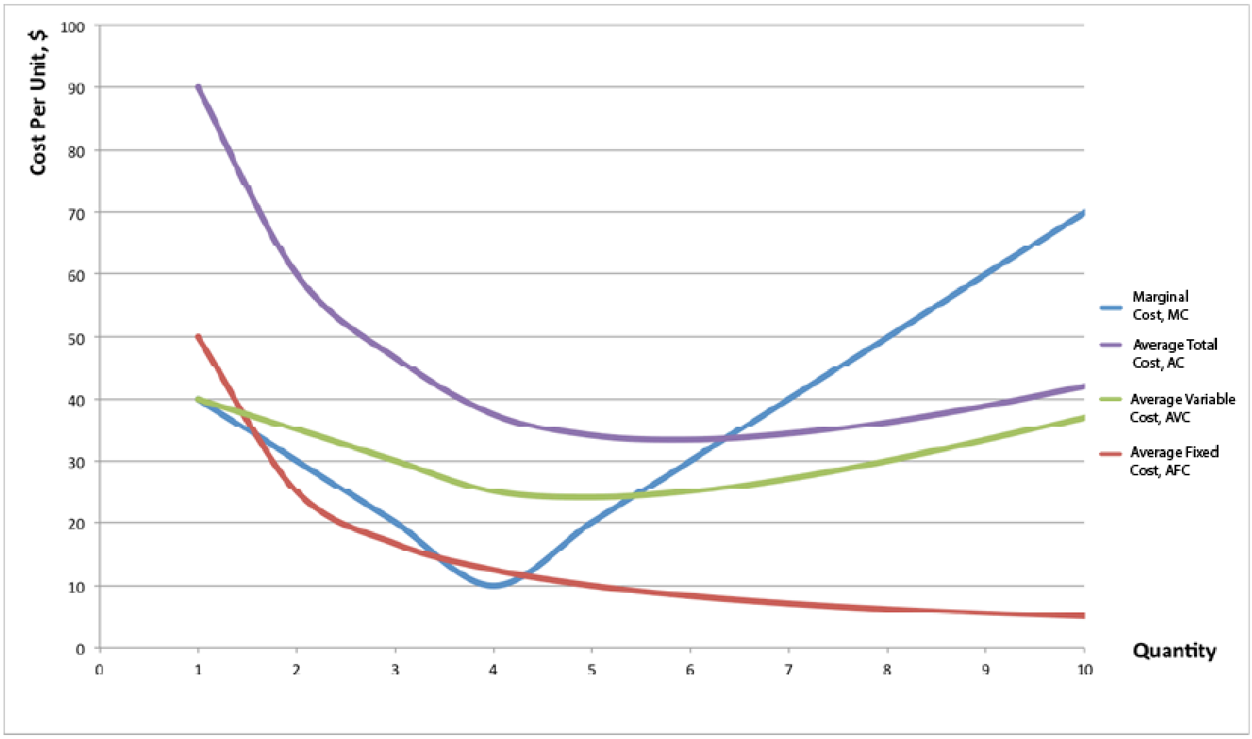


The Figure also shows variable and total cost curves. The variable cost increases with output, because extra output requires extra variable inputs. Probably the main variable input is the opportunity cost of the repairman's time. If the repairman is also the owner, this cost would not be written down on the books, but it certainly should affect the owner's decision. The owner might also have a hired helper who is paid an hourly wage which varies with time worked in a day. As we can see in the graph, the variable cost curve rises as output, Q, increases. At first, variable cost rises slowly, then it rises quickly. This is a pretty standard assumption made in these narratives. An alternative assumption often made for simplicity is that total variable cost VC is a line, so its slope doesn't change. Linear variable cost implies constant marginal cost, which as I mentioned earlier is an assumption sometimes made in order to narrow the focus of a story (we will do this later in the course when we study models of natural monopoly and duopoly).

Marginal, Average Fixed, Average Variable, and Average Total Cost Curves

The Figure below presents the four remaining short-run cost curves: marginal cost (MC), average fixed cost (AFC), average variable cost (AVC) and average total cost (AC).

Figure: Short-Run Marginal Cost, Average Fixed Cost, Average Variable Cost, and Average Total Cost Curves.



The marginal cost, average variable cost, and average total cost curves are derived from the total cost curves. They are all in the same units, “dollars (cost) per unit.” This is obvious for the average cost curves, but the marginal cost curve is as well, since it is the cost of the next unit, in dollars.[3]

From the Figure you can see that the marginal cost curve crosses the average variable cost curve at its minimum point. This is not by chance. Since the marginal cost indicates the extra cost incurred from the production of the next unit of output, if this cost is lower than the average, it must be bringing the average down. If this cost is higher than the average, it must pull the average up. Think of your own grade point average: if your average this term is higher than your overall average, your overall average will go up. If your average this term is lower than your overall average, your overall average will go down.

Long-Run Cost Curves

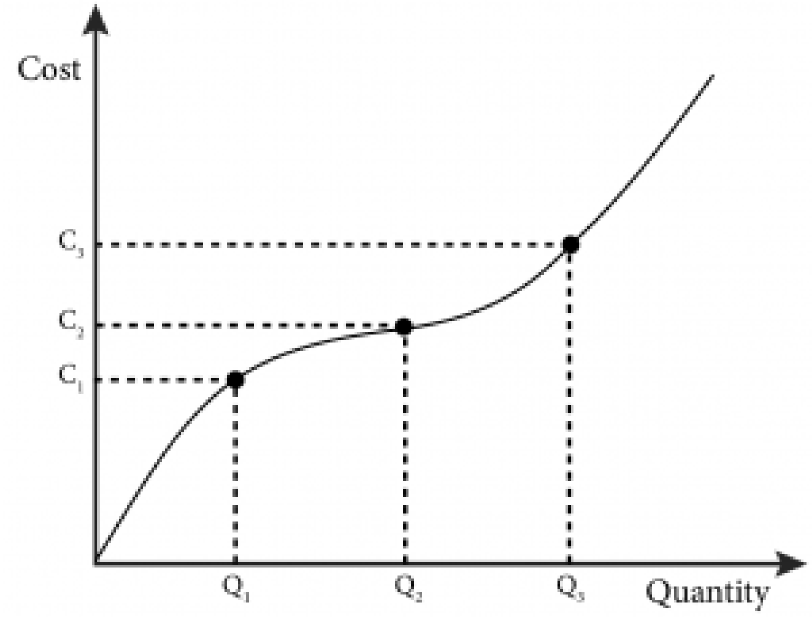
In the long run (1) all inputs are variable and (2) there are no fixed costs. There are not fixed-costs because in the long-run all inputs are variable. In the long-run, our air conditioning repair company can sell the truck and the tools. The restaurant can renegotiate or cancel its lease. And so forth. Thus, there are no fixed costs. In this section we look at the three long-run cost curves–total cost, average cost, and marginal cost.

Long-Run Total Cost Curve

The following figure shows a hypothetical long run total cost curve for a firm. In this case, for small quantities, total cost rises quickly – it is relatively expensive to expand production for a small firm. For moderate quantities, it rises slowly, so that the firm is producing extra units relatively cheaply. Eventually, with high enough production, the firm starts facing bottlenecks that leads costs to rise quickly again.

[3] Similar comments to earlier in the course about “smoothing out” the marginal curve apply here as well. In particular, the marginal cost curve given in this plot is per-unit changes in cost approximated by the slope of the tangent line to the total cost curve at each quantity produced. This is not that important, the only real value to doing this in the analysis is so that we can draw smooth curves rather than step functions.

Figure: The Long Run Total Cost Curve



Throughout the figure, the total cost curve is increasing: as we expand output we must use more inputs, and this results in increasing overall cost of production.

Long-Run Average and Marginal Cost Curves

A firm's **long-run average cost** (LRAC) is the cost per unit of output. In other words it is the total cost, LRTC, divided by output, Q:

LRAC(Q)=LRTC(Q)/Q.

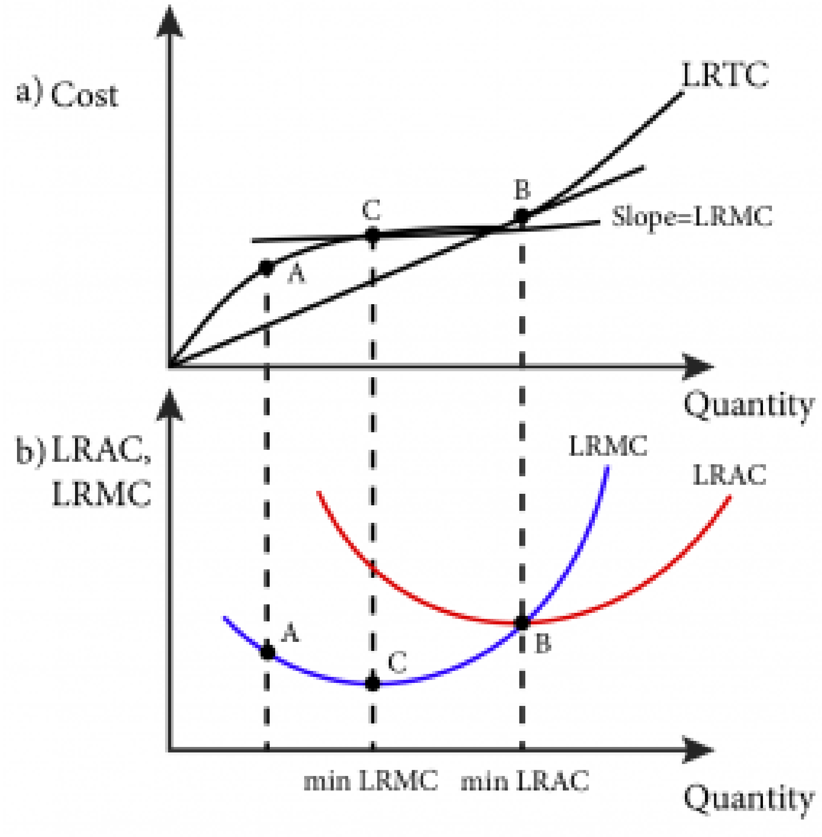
A firm's **long-run marginal cost** (LRMC) is the increase in total cost from an increase in an additional unit of output:

LRMC(Q)= ΔLRTC(Q)/ΔQ for ΔQ = 1

Where ΔLRTC(Q) is the change in total cost and ΔQ is the change in output. This is the same thing as the *slope* of the total cost curve *LRTC*(Q).

The Figure below illustrates the relationship between the total cost curve (panel a) and the average and marginal cost curves (panel b).

Figure: Deriving Long-Run Average and Marginal Cost Curves from the Long-Run Total Cost Curve



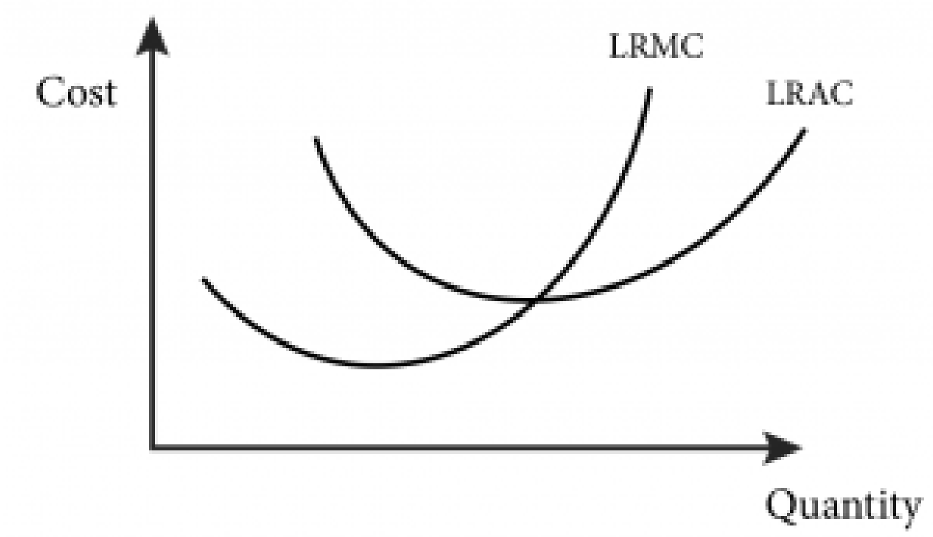
Note that at point B the marginal and the average costs are the same as seen by the intersection of the two curves in panel b, and average cost is at a minimum. This relationship between average and marginal costs is not a coincidence; it is always true. Exactly as in the short-run case, when average cost is above marginal cost, average cost must be decreasing. When average cost is lower than marginal cost, average cost must be increasing. And when average and marginal cost are equal, average cost is not changing, which means it is at a maximum or minimum – in this case, it's at a minimum. This relationship is described in the table and figure below.

I keep emphasizing this relationship between average cost and marginal cost because it is this fact that underlies the "perfectly competitive markets produce efficiently, i.e. at minimum long run average cost" argument that we will make later in this module. So make sure this fact makes sense to you.

Table: The Relationship between Long-Run Average and Marginal Costs

Relationship between AC and MC	**Resulting change in AC**
AC(Q)<MC(Q)	AC(Q) increasing
AC(Q)>MC(Q)	AC(Q) decreasing
AC(Q)=MC(Q)	AC(Q) constant

Figure: Relationship Between the Long-Run Average and Marginal Cost Curves

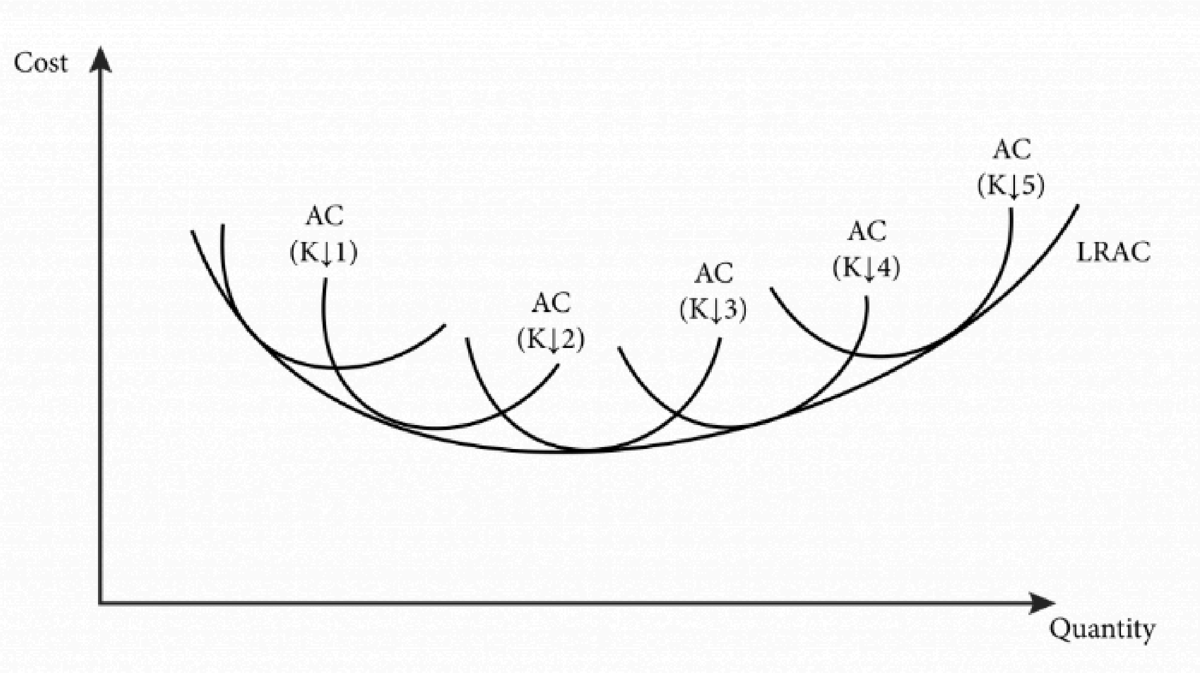


On the left half of the figure the average cost is above the marginal cost and thus the average cost is falling. On the right half the average cost is below the marginal cost and thus the average cost is rising. At the intersection of the two curves the average cost is at it minimum and the slope of the average cost curve is zero.

Short-Run Versus Long-Run Costs: The Advantage of Flexibility

Short-run average costs are constrained by the presence of at least one fixed input. *Therefore, in the long run the firm can always do at least as well as, and often better than, in the short run with respect to cost.* The long-run average cost curve is a type of lower boundary of the short-run cost curves. This can be understood most easily by thinking of a series of short-run average total cost curves, each one for a different level of the fixed input (e.g. capital) as shown below.

Figure: The Long-Run Average Cost Curve as the Lower Boundary of Short-Run Average Cost Curves.



The long-run average cost is essentially the same as picking the cheapest way to produce among the available short-run average cost curves. The long run average cost curve illustrates the benefit of flexibility: by being able to choose both inputs the firm can ensure the efficient mix of the inputs is being used at all times which keeps costs at their minimum points for all output levels. This flexibility means that we can expect that in the long run, the average cost of production is at least as low as, and generally lower than, in the short-run. This story suggests that, after a positive price shock occurs, e.g., oil trade embargo that reduces market supply, which makes firms' current production decisions suboptimal, time will to some extent cause costs to fall, supply to expand, and prices to fall.

So if you are the president and want to sanction Russian oil, it is better to do it early in your presidential term, and hope that there is enough time for other producers to switch to production processes that are cost-effective for higher quantities before the election.

Profit Maximization and Supply

We now study the supply side of markets: how firms' cost conditions define and affect their supply curves and the market supply curve. In our story, by summing up all producers' supply curves within a given industry, we can construct a market supply curve just as we constructed the market demand curve from individual demand curves. When we have an understanding of both market demand and market supply, we can then better understand the supply and demand model.

Output Decisions for Price Taking Firms

Before considering the production decisions of firms, we need to understand a few foundational assumptions to our story. First, we are going to focus on the behavior of *price-taking* firms. We talked about price-taking consumers earlier in the course, and price-taking firms are analogous. A firm is said to be a **price taker** when it has no ability to influence the price the market will pay for its product; it must take the *market price* as determined by the laws of supply and demand in a competitive market. A **perfectly competitive** market is a market in which there are many firms so that each individual firm's output has no impact on market equilibrium, output is identical across firms, firms have the same access to inputs and technology and consumers have perfect information about prices. All firms in a perfectly competitive market are price takers.

It is difficult to give an example real-world market that satisfies all the assumptions of a perfectly competitive market. This is intentional; the competitive market story makes strong assumptions in order to better be able to tell a coherent, straightforward, and hopefully enlightening tale. A common example given for a perfectly competitive market is the wholesale market for particular foods, such as say onions.[4] There are many onion farms in the US, each farm arguably takes the market price as given (onion farmers check what the market price is through websites and trade publications, and these reports are merely reporting the going prices), and each farm makes the decision as to the amount to produce based on the price and its own individual cost curve for production. In addition, the competitive market model implicitly assumes that the demand curve is composed of many individual consumers who take the market price as given. If there are enough separate purchasers of onions this might hold true for this market. Still, we should keep in mind that the perfectly competitive market story is just that—a story.

We assume each firm's goal is to maximize profit. A firm's **profit** (π) at produced quantity Q is the difference between its total revenue and its total cost:

$$\pi(Q)=TR(Q)-C(Q)$$

The total revenue is the quantity of the goods produced multiplied by the sales price of those goods.

$$TR(Q)=P\times Q$$

The total cost is the total cost curve C(Q) introduced earlier and represents the economic cost. The **economic cost** is another word for the opportunity cost, so it includes both explicit costs and implicit costs. This is distinct from the **accounting cost** which includes only the explicit costs, or those you would see in an accounting spreadsheet of the firm's costs. So profit is an expression of the economic profits of the firm, which considers economic costs. In economics, we focus exclusively on economic profits, because this is the relevant measure when it comes to decision-making. By now we have discussed many cases where the opportunity costs of an action, such as producing some quantity, are higher with what the accountants report as the explicit costs.

Let's now turn to the output choice of a profit maximizing, *price-taking* firm. The objective of maximizing profit means that firms must choose the

[4] The largest onion producer is 6% of the market, which is a fairly small share (sources: https://www.growingproduce.com/vegetables/americas-largest-onion-producer-changes-hands/ , https://www.onions-usa.org/all-about-onions/retail/us-production-and-availability/).

output level that maximizes the difference between total revenue and total profit. To determine this specific level of output the firm must ask how the production of one more unit of output contributes to both the total revenue and the total cost. For example, if a car manufacturer can produce one more car at a marginal cost of $15,500 and sell that car for a marginal revenue of $17,000, it knows that by producing the extra car its profits will increase by $1,500. Note that this is *not* the same as knowing that profits are positive because the calculation does not include fixed costs. All it means is that profits will be larger, or losses smaller, if that car is produced. Similarly if the additional car has a marginal cost of $18,000 to produce and can be sold for $17,000 then the manufacture and sale of this car would reduce profits by $1,000. So producing that additional car sounds like bad idea. If the marginal cost in this story sounds a lot like the firm's marginal willingness-to-accept (i.e. produce), that's because it's exactly that.

The term **marginal revenue** (**MR**), used in our example above, refers to the change in total revenue from a one-unit change in quantity produced. Marginal revenue is expressed

$$MR=\Delta TR/\Delta Q \text{ for } \Delta Q = 1$$

Here is the big thing that the assumption of price-taking buys us in our story: for a price-taking firm, the increase in revenue from the sale of an additional unit is always exactly the price of that unit. In other words, for a price-taking firm MR=P.

Any profit-maximizing firm would like to continue to increase output as long as marginal revenue is larger than marginal cost. A profit-maximizing firm would also like to reduce output as long as marginal revenue is lower than marginal cost. The incentive to increase or decrease output stops exactly when marginal revenue equals marginal cost. This is known as the *profit maximization rule*: profit is maximized when output is set where marginal revenue equals marginal cost.[5]

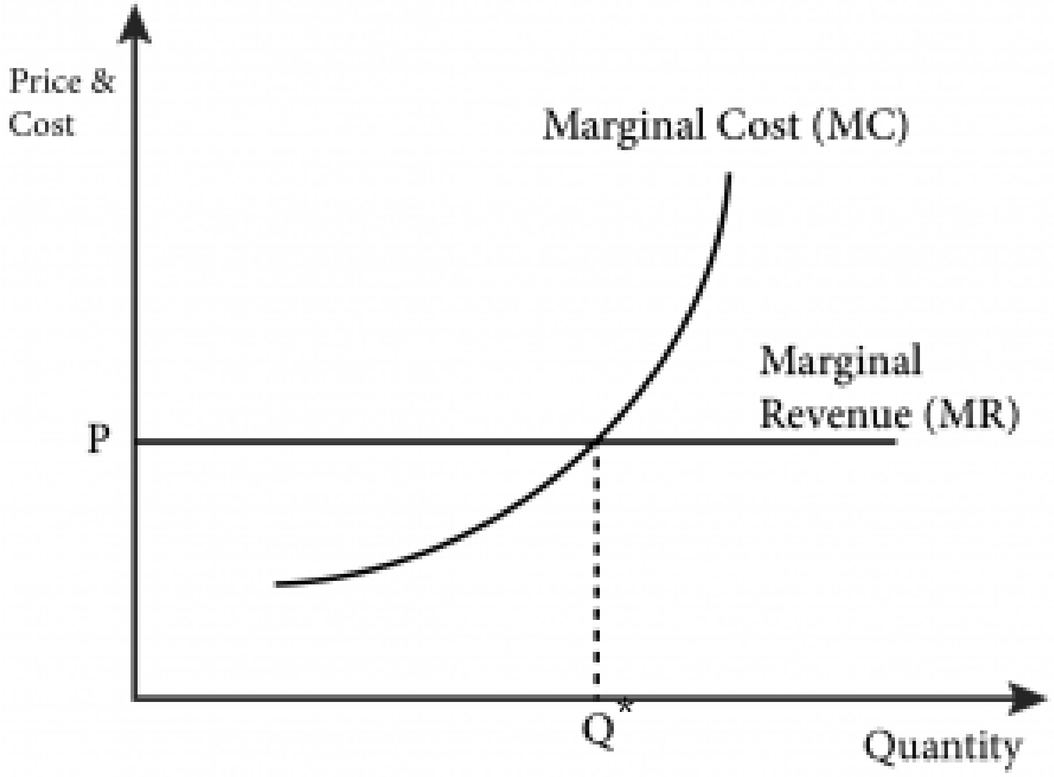
Earlier we learned that marginal cost (MC) is the additional cost incurred from the production of one more unit of output: MC=ΔC/ΔQ. Since the

[5] This is a necessary but not sufficient condition for profit-maximization. We also need marginal cost to be increasing at the quantity chosen.

profit maximizing rule stipulates that output should be set where marginal revenue equals marginal cost and since marginal revenue for a price-taking firm is the price of the good, we know that at the profit maximizing output level for the firm. That output is Q* that solves the equation P=MC(Q*).

The expression P=MC(Q*) gives us a relationship between the price, *P*, of a good and the quantity, Q*, that a profit-maximizing, price-taking firm will produce at that price. In other words, it gives us part of the individual firm's supply curve. I say only "part" because we still need to think about when the firm will shut down. But if the firm does in fact operate, then the marginal cost curve indeed is its supply curve. The Figure below illustrates this relationship.

Figure: Profit Maximization for a Price-Taking Competitive Firm



In the Figure we see that the firm's profit maximizing level of output is where marginal revenue equals marginal cost. For a price-taking firm, marginal revenue is equal to the price. So, as price increases the firm will increase production and when the price decreases the firm will decrease production (assuming throughout that the firm is operational). That sounds like a sensible prediction.

Short-Run Supply

To understand the short-run supply decision of the firm we have to be able to measure the firm's profits. The profit maximization rule, to set output such that marginal revenue equals marginal cost ensures that the firm is maximizing profit, but it does not ensure that the firm is *making* positive profits. In other words, following the MR=MC rule means that the firm is doing the best it can, which could be minimizing losses instead of making positive profits.

To go from the output decision to profits in our picture, we want to express everything in terms of dollars per unit. Ideally, profits will be a length or area or something like that in our figure. To do this, consider the following:

$\pi = TR - TC$

$TR = P \times Q$

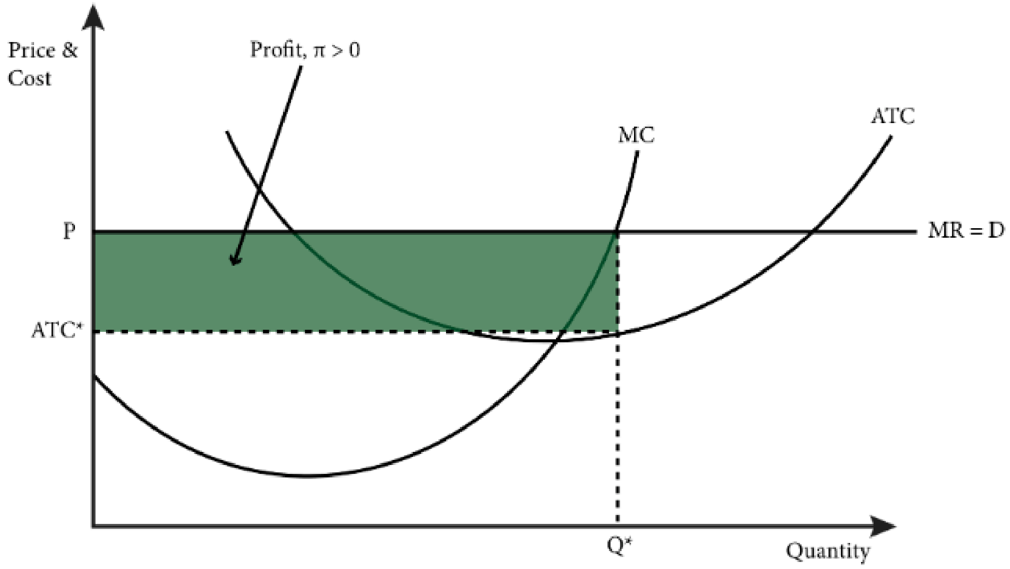
$TC = ATC \times Q$

Thus,

$\pi = (P \times Q) - (ATC \times Q) = (P - ATC) \times Q$

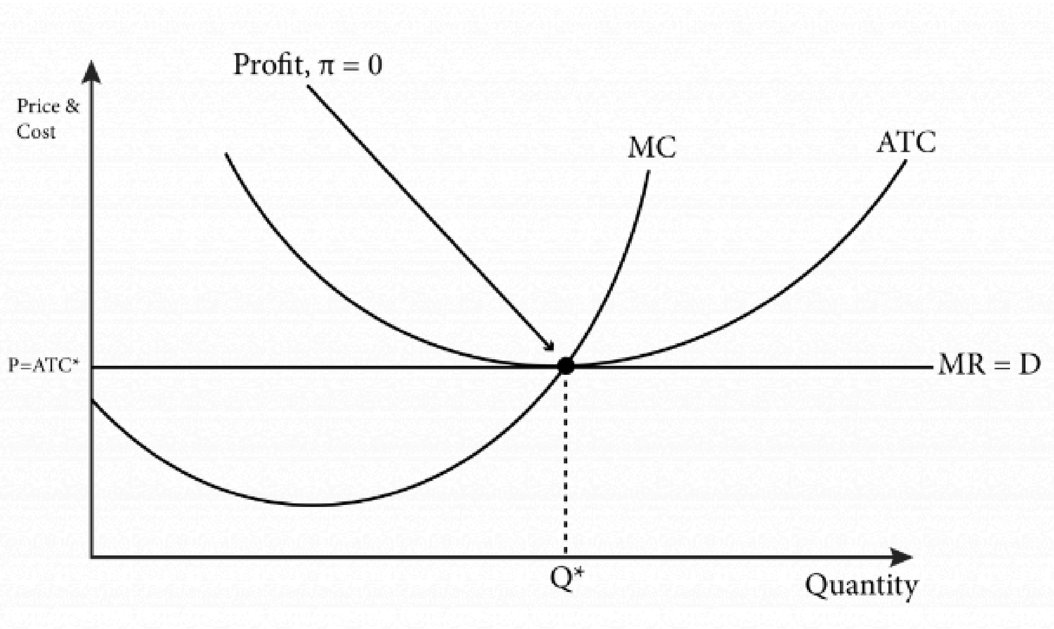
Since Q*, the quantity for which MR=MC, is always positive or zero, whether profit (π) is positive or negative depends on the price (P) relative to the average total cost at that particular Q*. Let's call that ATC*. If P>ATC* then π > 0 as seen in the Figure below. In the figure, profits (π) are represented graphically by the shaded area. Notice that the height of the rectangular shaded area is P-ATC, and the width is Q. Since the area of a rectangle is height times width, the area of this rectangle equals the profit.

Figure: Positive Profit: *P* > ATC*



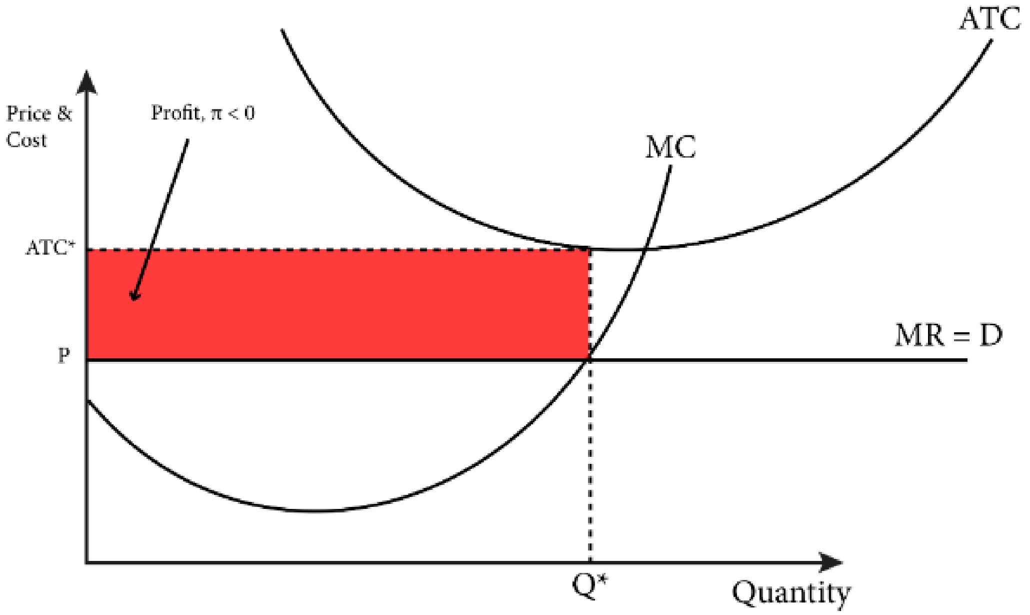
If P=ATC*, then π=0, as seen in the Figure below.

Figure: Zero Profit: P = ATC*



If P<ATC*, then π<0, as seen in the figure below.

Figure: Negative Profit (Loss): P < ATC*



It is tempting to think that if profit is negative, the firm should immediately shut-down or cease production of the good. But remember that in the short run, there are fixed inputs that cannot be adjusted immediately. For example, suppose you own a soap selling business requiring the leasing of a storefront. This lease is a three-month lease and the three monthly payments of $1000 each must be made regardless of whether the store is open or closed. Now suppose that the business is making enough revenue to cover all of the variable costs like the ingredients for the soap, the electricity bill, your own salary and *part* of the lease, perhaps $500 of the $1000. If you continue to operate the store, you will lose $500 per month – the part of the lease payment not covered by revenues. If you shut down you will lose $1000 per month until the lease runs out, because although there are no variable costs, there also is no revenue.

Consider the following alternative expression for profit:

π=TR-TC

=TR-(FC+VC)

=TR(Q)-FC-VC(Q)

The third line in this expression emphasizes that TR and VC are both functions of Q, and we know that if Q=0, then *TR*=0 and VC=0 as well. Shutting down means to set Q=0 so,

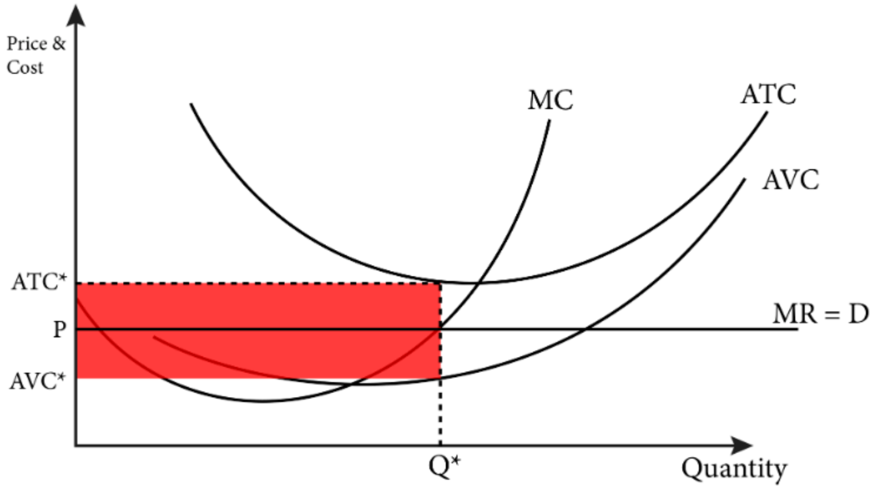
Level of profit if the firm shuts down:

π=0-FC-0=-FC

If we compare these two, we see that in the short-run, a firm should only shut down if the total revenue is lower than the variable cost, which is the same as the price being lower than the average variable cost.

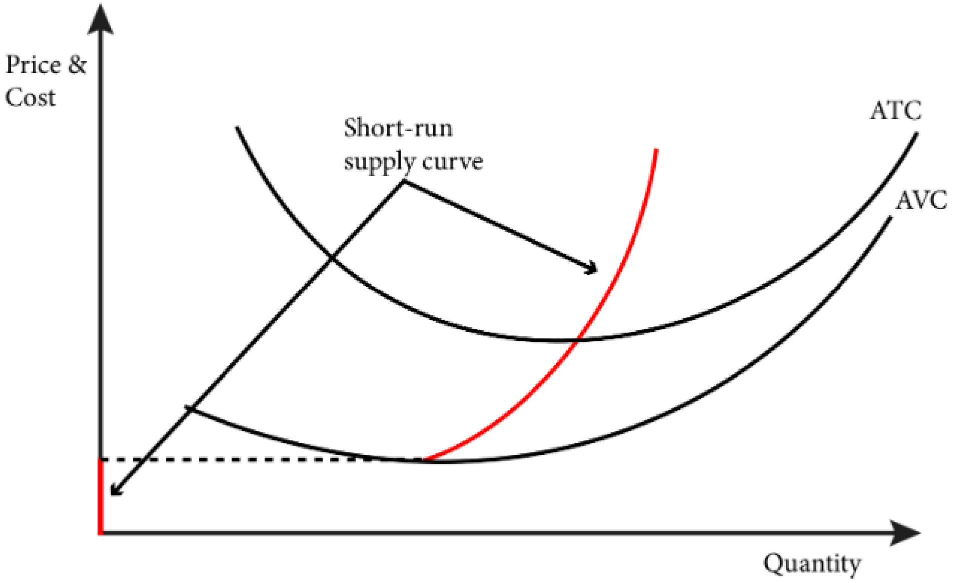
The Figure below adds the AVC curve in order to illustrate the situation where the firm is generating a negative profit but should continue to operate in the short run. At Q* the firm is making negative economic profits because (P-ATC*) is negative (recall π=(P-ATC*)×Q*). However, the firm is covering all of its variable costs (P>AVC*) and part of its fixed costs. So it should not shut down in the short-run. Of course, when the short run ends and the firm is able to adjust its previously fixed input, the firm should shut down if its total revenue is still lower than total cost. It should not pay its fixed inputs for the next year, since continuing to operate is not profitable.

Figure: Negative Profit but in the Short-Run Firm Should Continue to Operate



This understanding of the short-run output decision allows us to derive the firm's short-run supply curve. As long as the market price is above the firm's average variable cost, the firm will choose to produce output where P=MC. In other words, the firm's marginal cost curve above the AVC curve is the firm's supply curve. When price is below AVC the firm chooses not to produce any output at all so the supply for prices below AVC is zero. The figure below illustrates the firm's short run supply curve.

Figure: A Competitive i.e. Price-Taking Firm's Short-Run Supply Curve

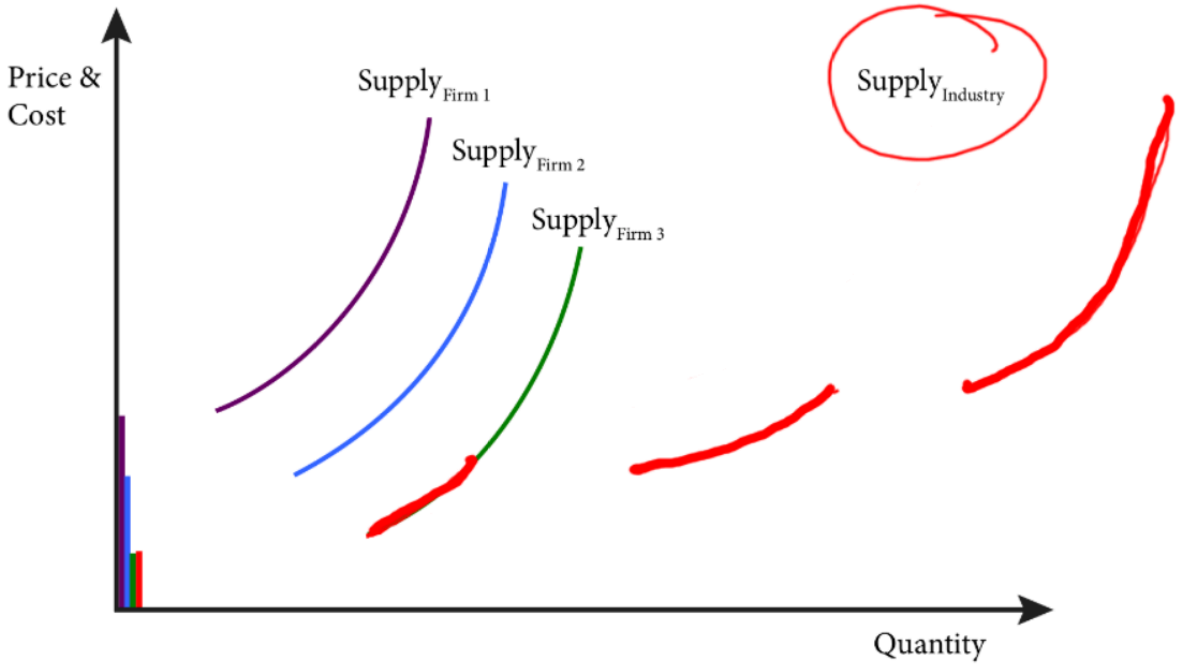


So the marginal cost curve gives the short-run supply curve, except for when the price is too low – lower than average variable cost – in which case the firm shuts down and produces 0. (You can think of the shutdown price for a firm as being analogous to the "choke price" for a demand curve.)

Every firm in a given industry has a short-run supply curve, but its precise shape depends on the firm's cost structure – the shape and location of its MC and AVC curves. Because each individual firm supplies a certain amount of output at every price, we can derive the industry supply curve by simply adding up these outputs across all firms in the industry. When we do this, we must take care to sum the quantities at every price, not the other way around.

The figure below shows how summing up all individual short-run supply curves yields the industry short-run supply curve, which represents the quantity supplied in the short run at every price.

Figure: Deriving an Industry Short-Run Supply Curve



This industry short-run supply curve looks a bit weird with three firms because when firms switch from being short-run-shut-downed to operating (as we think of the price rising), this leads to a discontinuous jump in quantity supplied. With a large number of firms (which is assumed in this model), the industry supply curve will not jump like this, it will instead be continuous – which is how we usually draw it.

Long-Run Supply and Market Equilibrium

In our short-run supply story, we made a lot of assumptions. One assumption that was made but perhaps not highlighted enough was the fact that in the short-run, new firms cannot enter the market. The reason this was made is simple: in the short-run, the fixed inputs (corresponding to fixed costs) cannot be adjusted. Thus, the firms not in the market do not have the fixed inputs needed to operate in the market. So they simply cannot produce. Therefore, the short-run supply curve consists only of the firms currently in the market. Under the story we told above, the firms all have rapidly increasing marginal costs at some relatively small quantity of output; therefore, each firm can only supply a relatively small part of the market. This story is internally consistent because if each firm only supplies a relatively small part of the market, then each firm can reasonably be assumed to be a price-taker, which was foundational to the perfectly competitive market model we are analyzing. Together, this story implies that the short-run supply curve is upward sloping, as we drew it earlier.

In the long run, things are more complicated. In the long-run, firms do not have any fixed costs; all production costs are variable. So a firm's profitability is determined solely by the long-run average total cost curve. A profit maximizing firm still sets output such that marginal revenue equals marginal cost, and since marginal revenue for a perfectly competitive firm is equal to the market price, the long-run marginal cost curve above the long-run average total cost curve (LRATC) represents the firm's supply curve.

As in the short-run case, the firm's profits depend on the price relative to the long-run average total cost at the optimal output level for the firm, Q*. In the Figure below, the price is above LRATC so the firm is making positive profits. As long as profits are not negative the firm will continue to produce. So the portion of the long-run marginal cost (LRMC) that lies above the LRATC is the firm's supply curve. This is shown in the next figure.

Figure: Positive Profits in the Long Run

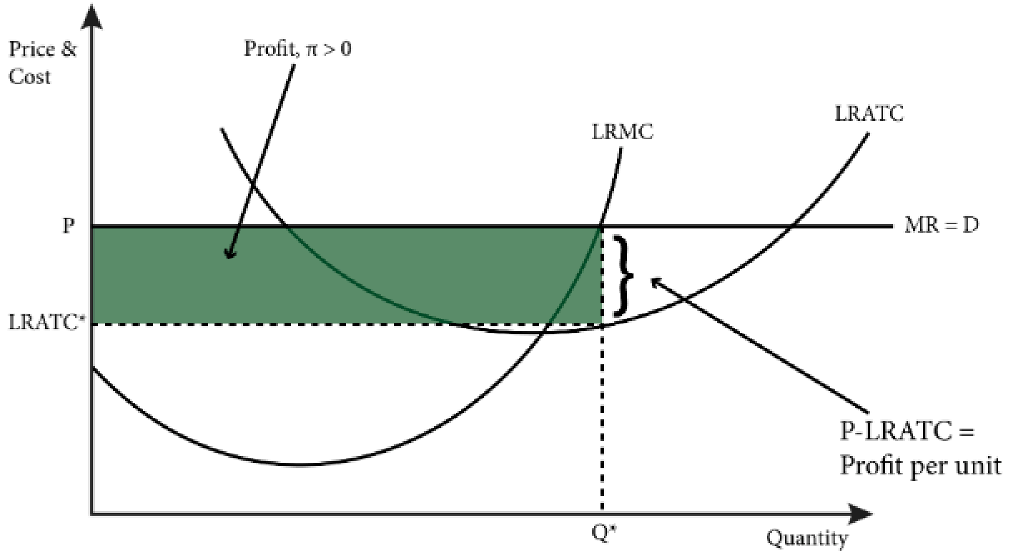
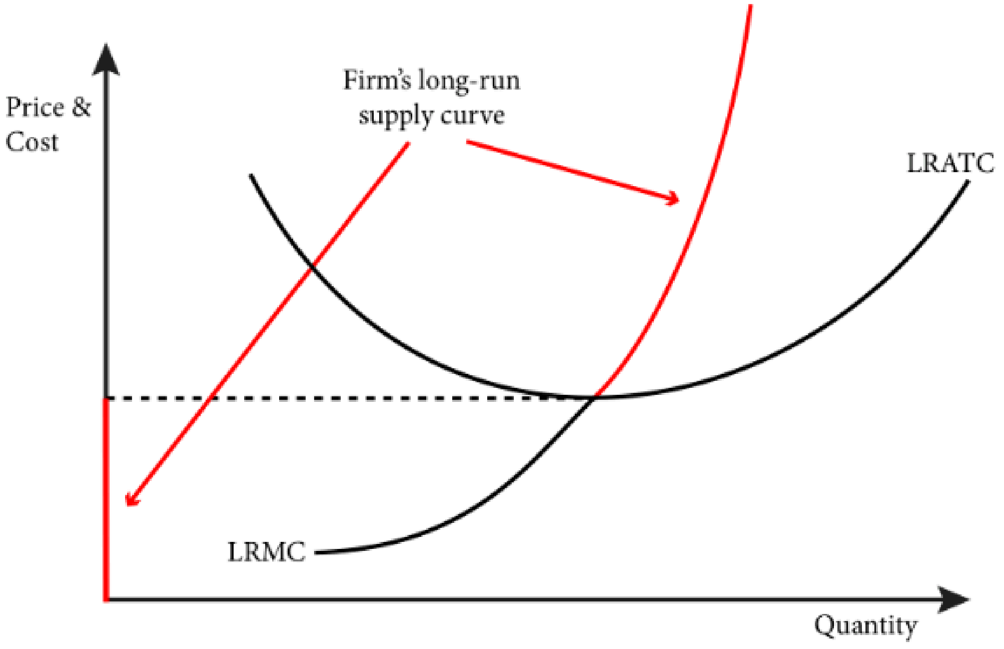


Figure: The Long–Run Supply Curve of a Perfectly Competitive Firm



That all seems like a straightforward copy & paste from our earlier discussion. The wrinkle, however, is that in the above figure, we assumed that the firm was making positive profits in the long-run. But if so, wouldn't new firms enter the market to try to capture these profits?

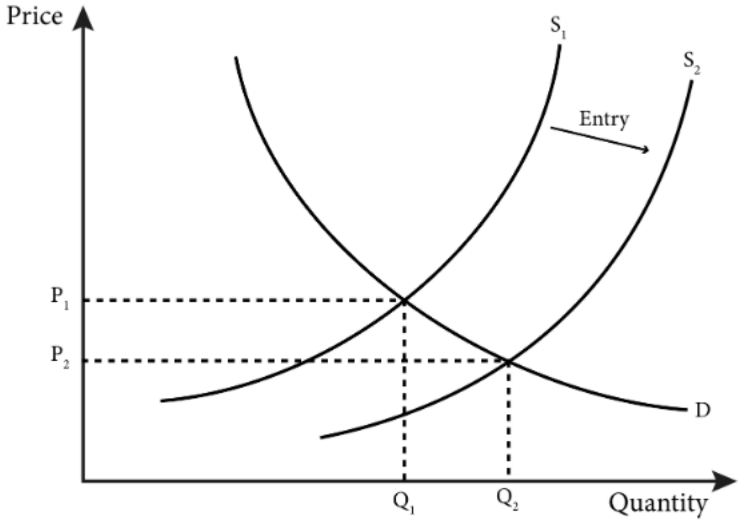
Free Entry and Exits with Homogeneous Firms

To derive the long-run *market* supply curve, we have to think about how firms enter and exit industries in the long run. The perfectly competitive market model assumes the industry has **free entry** and **free exit:** there are no special costs, such as technical or legal barriers, to firms entering and exiting the industry. Note that "free" does not mean there are no costs: there will be monetary costs (the fixed costs of capital). But if a firm not currently in the market can pay those monetary costs, then it can enter the market. This means the calculus of whether a firm enters hinges entirely on money, not on something nonmonetary (like a government license to operate). This free entry/exit assumption is critical to the perfect competition model. Barriers that block firms from entering or exiting will create an environment that has only limited competition, and this model is unlikely to be a good way for us to think about those markets. We will talk about such barriers later, when we examine models of monopoly and duopoly.

If there is free entry and exit, the next question to answer is when will firms choose to enter and exit a market? To answer this question, we will assume for now that all firms in a market are *homogeneous*. That is, they have identical technologies and cost structures, or more simply they all have the same LRATC and LRMC curves. This assumption substantially simplifies the story. We will examine firms that have heterogeneous costs in the next module.

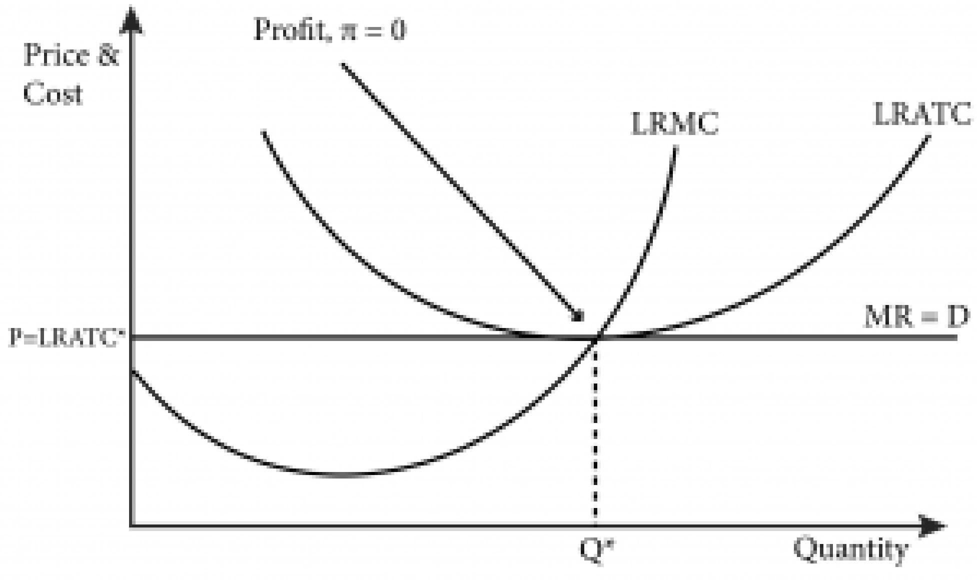
The previous figure illustrated the case where the market price is such that the homogenous firms are making positive profits. Positive profits in this case mean that the firm is getting better than normal returns, or that this is an exceptionally profitable market to be in. Other firms, not currently in the market, will see these profits and decide that this is a good market to enter. When new firms enter the market, they add their output to the total supply and the market supply increases.

Figure: New Entrants Increase Market Supply and Lower Equilibrium Price



As new firms, drawn by positive profits, enter the market, the added supply lowers the **equilibrium price**—the price at which quantity demanded equals the quantity supplied. This lowering of the equilibrium price lowers all firms' profits. Entry will continue to occur as long as firms' profits are positive, and so this process will continue until the equilibrium price has reached the point where the LRATC and the LRMC cross, or where there are *zero profits* as shown in the figure below.

Figure: Equilibrium With Zero Profits



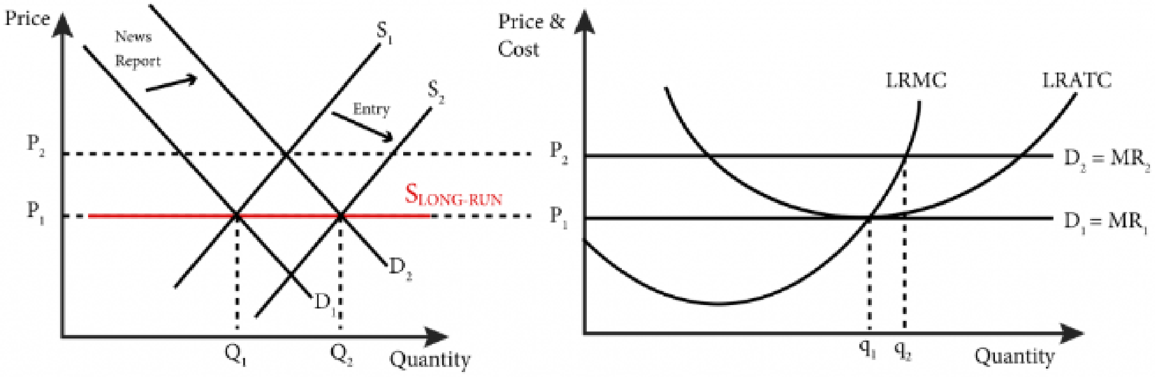
The market exit dynamics work similarly to the market entry dynamics. Firms that are currently producing and supplying their output to a market where the equilibrium price is below their LRATC are making negative profits, which by the definition of opportunity cost means that other opportunities exist that yield higher returns. This fact will cause existing firm to exit the market, lowering the market supply curve, and raising the price. This will continue until, again, the firms are earning zero profits.

Long-run market equilibrium actually includes two conditions:

- a market clearing condition where the price equates the quantity supplied to the quantity demanded;
- a zero profit condition, since only if existing firms earn zero profits does the market not attract any new entrants nor lead to any exits

With homogenous firms and free entry, the long-run market supply curve is simple, yet intriguing: the price has to equal the minimum long run average total cost (which is a particular number, the same for all the identical firms). This is illustrated below. Furthermore, notice that the long-run supply curve in this model is *perfectly elastic* ! Indeed, if the cost structure is such that additional identical firms can always be added to produce more output, then the long-run supply must be perfectly elastic, despite the short-run supply being upward sloping (because the little incumbent firms cannot adjust production except at increasing marginal cost). This is an extreme case of long-run market cost adjustment.

Figure: The Long-Run Market Supply Curve



The figure above also illustrates what happens in the short vs long run from an increase in demand (in the left-hand side of the figure). Due to a News Report, demand for the product increases from D to D_1. The short-run supply curve S_1 is upward-sloping, because the current firms in the market are producing using cost structures with increasing marginal cost. The price then rises in the short-run, and there is a short-run market equilibrium with a higher price and quantity. In the long-run, firms enter the market. Earlier in the course, firm entry was represented by the supply curve shifting rightward. Here we can represent that by the short-run supply curve shifting rightward, as shown in the picture. It shifts rightward until the price falls back down to the minimum long run average total cost (which again is just a single number since these firms are all the same). Thus the long-run supply curve is horizontal.

It's an interesting story, but that assumption of infinitely many potential homogeneous firms that can enter is a big one. We will consider the case with heterogeneous firms in the next module.

Perfect Competition and Efficiency

Let us now turn to the efficiency properties of the perfectly competitive market model with homogeneous firms. A while ago, we studied the concepts of consumer and producer surplus and defined Pareto efficiency and Kaldor-Hicks efficiency in the context of the supply and demand model. We discussed a story about how prices adjust to conditions of excess supply and excess demand until a price that equates supply and demand is reached. What this means is that the market ensures that everyone who values the good more than or equal to the marginal cost of producing it (the marginal willingness to accept) will find a seller willing to sell the good and that all opportunities for consumer and producer surplus will be exploited. This is how we know that total surplus is maximized, i.e., the outcome is efficient in both the Pareto and, more importantly, the Kaldor-Hicks sense.

In the perfectly competitive market model, neither consumers nor producers have any individual influence over the market price, although their actions taken collectively determine the market price. As we have discussed, in this model, each individual active producer produces output until marginal cost of production equals the price. Meanwhile, individual consumers purchase the good if their marginal benefits from the purchase are greater than the price. Then, if the market price is too high, firms produce more than demanded and the resulting surpluses lead to a fall in the price; if the price is too low, firms produce less than demanded, and the excess demand leads to a rise in the price. This intuitive story suggests there is a force for the price to be exactly at the point where the quantity supplied and demanded are equal, i.e. the supply and demand curves cross. If that crossing point is the market price, then, in a perfectly competitive market, all the transactions that increase surplus, i.e. all the transactions where the benefits to the purchaser exceed the costs to the producer, will take place (we saw this in the original supply and demand graph). We say that this outcome is efficient, and total surplus is maximized.

Since we have talked about the short-run and long-run at length in our story, at this point, it is worth discussing two kinds of efficiency. So far when we talked about efficiency, we were referring to **allocative efficiency**, also called **efficiency in allocation**. Efficiency in allocation means that given current incumbent firm cost structures (i.e. current willingnesses-to-accept or -produce) all trades take place where the benefit of the transaction exceeds the cost: at the market equilibrium, there are no sellers and buyers out there who are being shut out of the market such that the sellers' marginal cost is less than the buyers' marginal benefit (if there were such sellers and buyers shut out, that would be allocatively inefficient since there are positive surplus deals left unmade). The perfectly competitive market model is always allocatively efficient, in the short or long run, since a single price rations the good such that the marginal benefit of the last unit sold is equal to the marginal cost of the last unit sold. This is the same as how market equilibrium maximizes surplus in our first discussion of the supply and demand model, where the equilibrium price equated the marginal willingness to pay (demand curve) with the marginal willingness to accept (supply curve).

Economists also talk about **productive efficiency,** or **efficiency in production**, which means that the firms are producing the good at the minimum possible average cost per unit.

To further contrast these two notions of efficiency, a situation where only high-cost producers are in the market because, for some reason, low-cost producers are being shut out, and yet conditional on the producers that are in the market all mutually beneficial transactions take place, would be allocatively efficient – it is efficient in the sense of maximizing total available surplus *given* the supply and demand curves of the market. But it would not be productively efficient, since the market supply curve would be inefficiently high, i.e. supply would be inefficiently restricted, because the only firms in the market are high cost.

With homogeneous firms, the competitive market model is efficient in production in the long-run, after entry/exit has taken place and economic profits are driven to zero. In the short-run, firms can produce at output levels that differ from minimum long-run average total cost, but in the long-run, with free entry/exit, the price must equal the minimum long-run average total cost (which with homogenous firms is a single number, the same for all firms) and so firms must be producing efficiently (otherwise there would be more entry/exit). The fact that firms choose output at marginal cost, marginal cost crosses the average total cost curve at the latter's minimum, and that free entry/exit require long-run equilibrium price to equal long-run average total cost (for zero profits), are together needed to imply this productive efficiency result. (With heterogeneous cost firms, the story is more complicated. We will discuss that in the next module.)

This might all seem somewhat tautological at this point, so it is worth reminding ourselves what happens when an assumption fails. A simple example at this point is when a government policy distorts the market. Consider a specific tax on producers, so producers have to pay something to the government for each unit of good produced. Then producers no longer produce at the quantity where price equals marginal cost, they instead produce where price equals marginal cost + the tax payment (because to the producers, the tax payment looks like an added variable cost). This means each producer's supply curve shifts up, so the aggregate supply curve shifts up (in the long-run the aggregate supply curve is a horizontal line under the homogenous firms assumption). We can then draw in the deadweight loss due to the transactions that do not take place because of the specific tax, as we did at the beginning of the course. The result is not allocatively efficient, since some surplus-creating transactions are squeezed out of the market by the tax. Moreover, because each firm now effectively acts based on the curve Marginal Cost + Specific Tax, this new curve no longer crosses each firm's long run average total cost curve at its minimum. Each firm produces an inefficiently small amount of the good, so we lose efficiency in production as well.

Economists often argue that "free markets" are efficient. By "free" we mean that prices freely adjust, and that there are no institutional or competitive controls that prevent prices and suppliers from adjusting to equilibrate the market until the efficient outcome is achieved. By efficient we mean both allocatively efficient i.e. that there is no dead weight loss, and efficient in production i.e. minimum average cost production. In this module, we have studied a narrative that is together called the "perfectly competitive market model," and which suggests that under some quite strong conditions the story of free market efficiency is true. These conditions include: there must be many buyers and sellers, and the sellers must have cost structures that have increasing marginal cost for relatively small quantities (so that there are many small sellers in equilibrium and none individually can manipulate the price); the good must be homogenous; there must be free entry and exit; there must be complete information about the good and prices on the part of buyers; and there must be no transactions costs. If this sounds like a lot of assumptions, that's because it is. Later in the course, we will analyze models where we don't make these assumptions. In some cases, those stories will reveal inefficiency if markets are permitted to operate freely. And in some cases, government can do something to correct that. In other cases, we will surprisingly get efficient outcomes even if the market is not perfectly competitive, or we will get inefficient outcomes that government cannot plausibly correct. Of course, the real world is far more complicated than any of these stories, and the perfectly competitive market model is perhaps one of the most fanciful stories economists have told. Nevertheless, it is valuable for us to understand this model, since it provides one of the most common stories that economists have in mind when they say, "free markets are good."

License

Unless otherwise noted, material is derived from the following sources. Note that the material from the first source has been changed substantially from the original version.

Patrick M. Emerson. *Intermediate Microeconomics.* *https://open.oregonstate.education/intermediatemicroeconomics/chapter/module-1/*

Material at the beginning of the document has been copied verbatim from the following source

> Ariel Rubinstein. *Economic Fables.* Cambridge, UK: Open Book Publishers, 2012, https://doi.org/10.11647/OBP.0020