Module 6: Monopoly

This module begins our examination of market power and models of competition when there is only one large firm, or only a few firms, serving a market. In this module, we will focus on *monopoly* (one firm). In the next, we will examine duopoly (two firms), and on the way to that, we will learn a bit of Game Theory, which has wide applicability in microeconomic-style modeling in many disciplines (including public policy and political science).

Monopoly

A **Monopoly** is the sole supplier of a good for which there does not exist a close substitute. The characteristics of monopoly markets are a single firm, a unique product and barriers to entry. But what are those barriers to entry? Two main sources of barriers include cost conditions that make it cheaper for one firm to produce than for many, and government regulations that create legal barriers to entry that prevent other firms from competing in the same market.

Cost advantages can come from a number of sources. A firm might control an essential input, for example one company might control the only source of a rare earth element use in high performance batteries for automobiles. A firm might have a better technology or method of producing a good or service that gives it a cost advantage. Of course this advantage only lasts as long as the firm is the only one with the technology or method. For this reason, many firms keep their production processes a closely guarded secret.

Another source of monopoly are very large fixed costs associated with production. If the fixed costs of production are large enough relative to the demand for a product, it may be true that only one firm can make non-negative economic profits. This is referred to as a **natural monopoly**: when one firm can supply the market more cheaply than two or more firms. For example suppose an entrepreneur wants to start a concrete pumping service in her isolated small town to provide poured concrete to difficult to reach places that normal concrete trucks are unable to pull up to. Concrete pumping trucks are very large and very expensive. The town might have enough demand to justify the cost of one such truck, the entrepreneur will attract enough business to cover the cost of the truck itself and the other accounting and opportunity costs of the business. But there might not be enough demand to justify the purchase of a second concrete pumping truck if another business decided to start up. Thus the monopoly that results is a natural result of the large fixed cost relative to the limited demand.

Using familiar curves, the following figure, reproduced from Friedman (2002), illustrates the difference between the cost structure of a natural monopoly (panel a) with that of the (homogenous firms) perfectly competitive market model (panel b). In a natural monopoly, one firm, with average cost curve AC, has declining average cost to and through the efficient (surplus-maximizing) level of output Q_E. By contrast, in the perfectly competitive market model (panel b), each firm is a small share of the market and produces at minimum average cost, and average cost is most definitely *increasing* in quantity produced if one firm were to hypothetically try to serve the efficient quantity Q_E. In a natural monopoly, it is efficient based on production cost considerations alone (ignoring for the moment the allocative efficiency problems monopoly entails, which will be discussed at great length shortly) for one firm only to produce the total output. That is, one firm is the only way to achieve efficiency in production. In the perfectly competitive market model, it is efficient based on cost considerations for many firms to each produce a small amount of the output (then, as the story goes, each firm individually takes the price as given, and the total output produced and sold is both productively and allocatively efficient, as we have discussed).

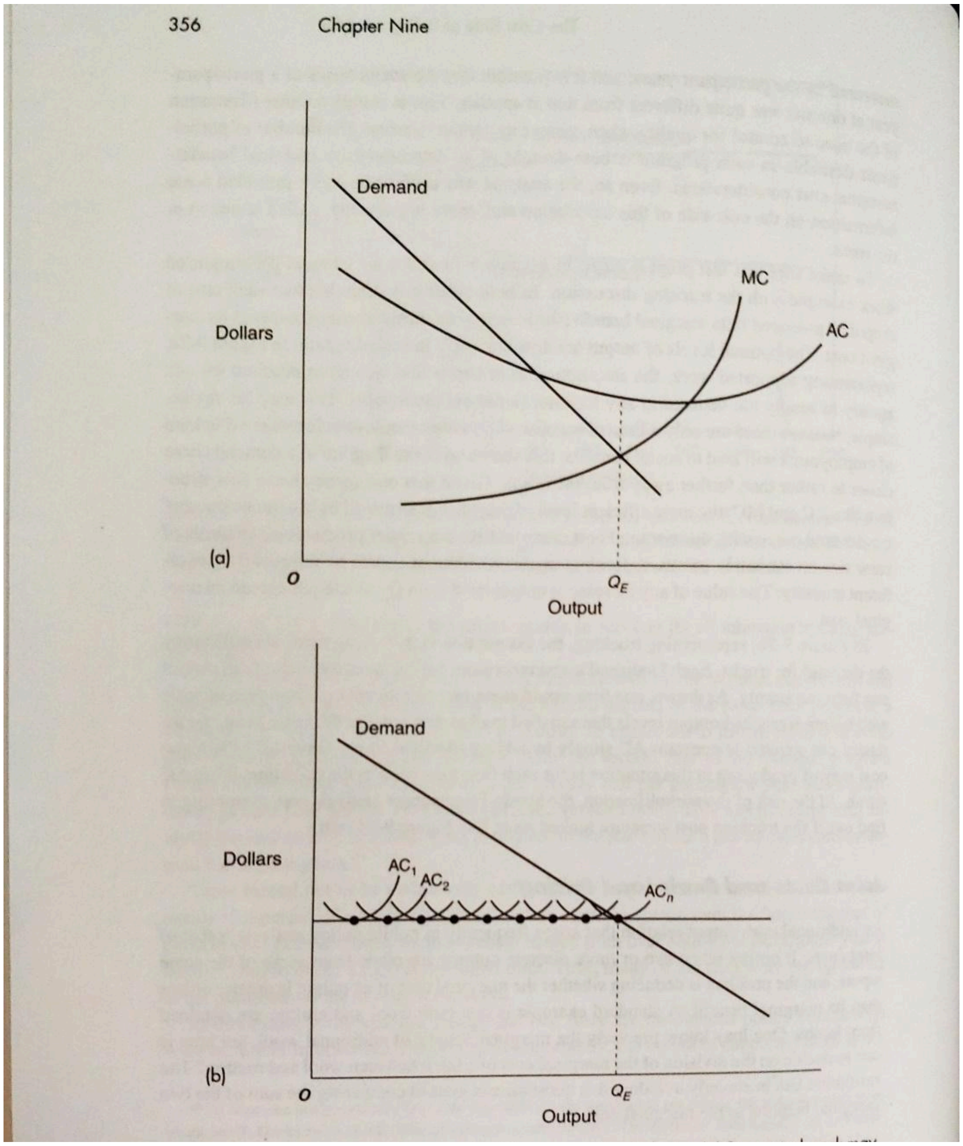


Figure: Natural Monopoly (a) versus perfectly competitive market (b). In figure (b), the average cost curve of the *n*th firm is shifted over by the quantity produced of all firms prior to it in the ordering; each firm produces at minimum AC, so the total quantity produced is the market equilibrium Q_E. The long-run supply curve in panel (b) is horizontal (because we are assuming infinitely many homogenous little firms can enter).

Another important source of monopoly power is government. The government itself owns and manages monopolies, for example the federal government runs the U.S. Postal Service, and local governments are often the sole provider of utilities such as water and sewer services. The government also provides protection from competition for individual firms, guaranteeing their monopoly status. The most common protection the government offers is patent protection, but government can also require licenses to operate businesses, for example to a local hospital, and issue only one per market or they can grant a firm the right to operate as a monopoly, for example a municipality might grant one company the right to collect residential trash.

A **patent** is an exclusive right to an invention, which excludes others from making, using, selling or importing it into the country for a limited time. In other words, the inventor is granted protection from any direct competition for a time period, generally 20 years from the time an application is filed, and in exchange the invention becomes public knowledge when the patent is granted. Why does the government offer patent protection? As we will see in the next section, being a monopolist often allows companies to collect positive economic profits – a better than normal return on investment. So patents give companies a reward for investing in new technologies, new drugs, and new methods that are valuable to society in order to incentivize them into investing in such discoveries.

Profit Maximization for a Simple Monopolist

All profit maximizing firms, regardless of the structure of the markets in which they sell, maximize profits by setting output such that *marginal revenue* equals marginal cost. What is different in the case of monopoly is the **marginal revenue curve**. Perfectly competitive firms take prices as given, their output decisions do not change market prices and so the marginal revenue for a perfectly competitive firm is simply the market price: in that model, the revenue for an additional sale is the price, and one firm selling more or less does not affect the price (since each firm is tiny). Monopolists, on the other hand, are ‘price makers’ in the sense that they can set their price by picking a point on the demand curve. Note that they are not free of constraint; the demand curve dictates the maximum price they can charge for every quantity level.[1] The task for the profit-maximizing monopolist is to determine which point on the demand curve maximizes their profits. A downward sloping demand curve will mean that the firm faces a trade-off: they can sell more but must lower their price to do so. This means that the marginal revenue of a monopolist will depend on their output decision.

The total revenue TR for a firm is the amount of goods they sell, Q, multiplied by the price at which they sell the goods, p. Note that Q is both the firm's output and the market output as there is only the one firm supplying the market. Moreover, note that we are assuming the firm posts only one price p. The assumption that the monopolist can only post one price, and all consumers are permitted to buy at that price, is why we call this a model of a **simple monopolist.**

$$TR = pQ$$

The firm's marginal revenue MR is the change in total revenue from the sale of one more unit of their output. Thus a firm that earns ΔTR extra revenue from the sale of ΔQ more units has a marginal revenue of:

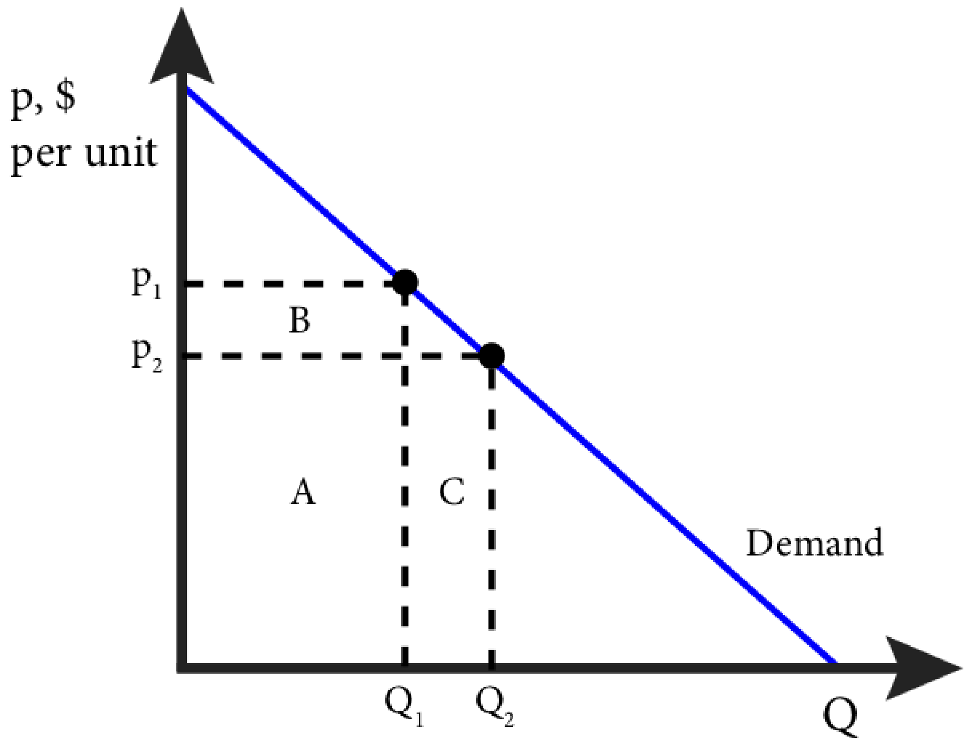
$$MR = \frac{\Delta TR}{\Delta Q}$$

In the perfectly competitive model, this was simply the market price. But because of the price-quantity trade off from the demand curve, the marginal revenue of a monopolist will depend on its output. With downward sloping demand, a monopolist who wants to sell one more unit must lower its price to do so. Thus, a monopolist's marginal revenue is a constantly declining function. It also declines at a rate greater than the demand curve because to sell more the monopolist must lower the price

[1] The fact that the monopolist faces a downward sloping demand curve, rather than a vertical demand curve, means that customers will, in fact, go elsewhere when the monopolist raises its price. This means there is some substitutes out there (or customers have income constraints; in the morbid case of a life-saving drug, for example). The first key assumption, then, in the story of the monopoly, is that the monopolist firm we are studying is large enough in terms of sales that when it chooses to change the amount it sells, the market price changes ("enough for it to matter"). Equivalently, when the monopolist changes the price it charges, the amount it sells changes smoothly, rather than going to zero if the monopolist charges above market price or "practically infinity" if the monopolist charges below market price (as is what we assume in the perfectly competitive model). The second key assumption is that there is no other firm in the market worth modeling strategically. We'll revisit this second assumption in the next module.

of all its goods. The Figure below demonstrates the marginal revenue trade off for a linear demand curve.

Figure: Monopoly, Revenue and Marginal Revenue



At price p_1 the total revenue is p_1Q_1, which is represented by the areas A+B. At price p_2 the total revenue is p_2Q_2, which is represented by the areas A+C. Area A is the same for both so the marginal revenue is the difference between B and C or C-B.

If we naively and wrongly assumed that the marginal revenue curve were the demand curve, then from the change shown in the figure, the firm would have to *gain* revenue equal to approximately area C, because (under the wrong assumption) the extra revenue from selling the units could be read off the demand curve. Clearly this is false. By lowering the price, the firm loses area B, and gains only C-B. Thus the marginal revenue from the greater output must be less than the demand curve. In general, as quantity increases, the marginal revenue will be increasingly lower than the demand curve. For example, for the first unit sold, the marginal revenue is the demand curve (the area B does not exist), because the monopolist does not have any “already sold units” that it has to lower the price on, in order to sell the next unit. But if the monopolist is already selling a lot of units, then when it lowers the price it posts, it will lose a ton on the units it was “already going to sell;” that is, area B becomes increasingly large relative to area C.

To give a concrete algebraic example that we will use very often, suppose the demand curve were linear. So $P = a - bQ$ for numbers a, b. Then total revenue is

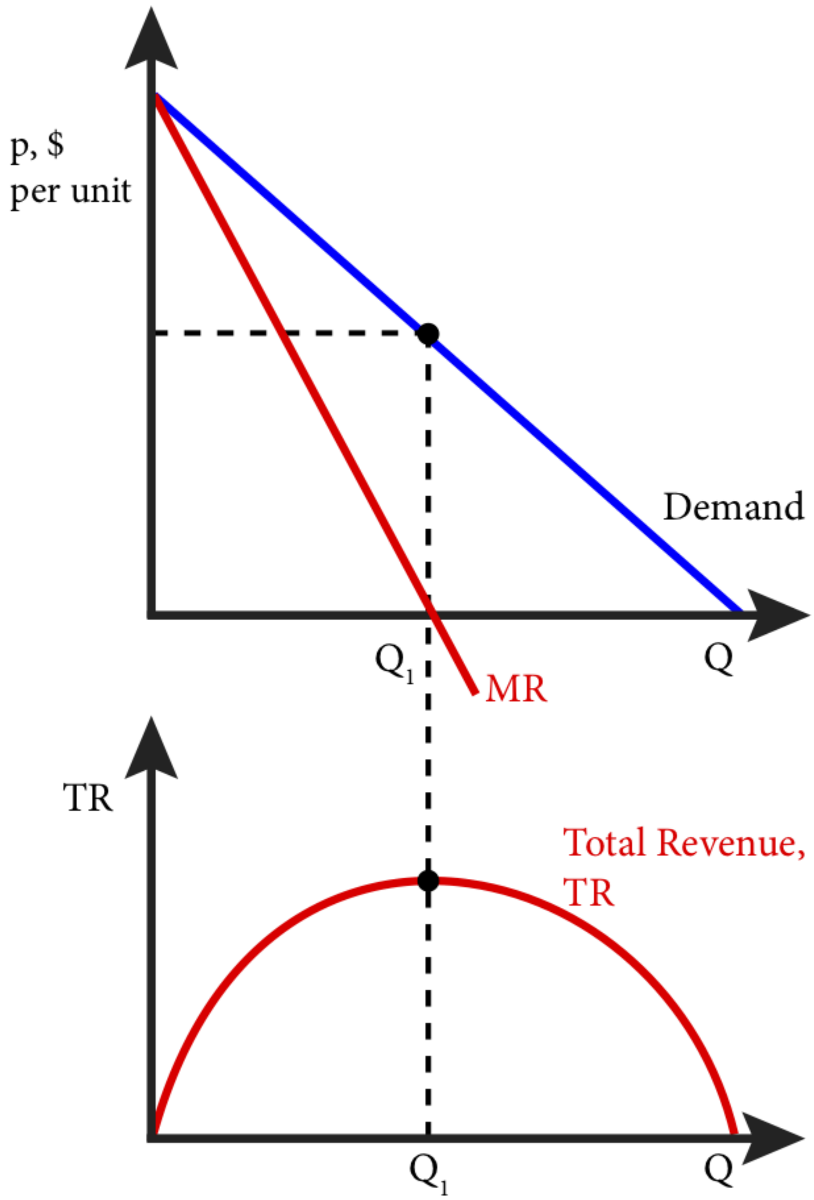
$$TR = P \times Q = (a - bQ) \times Q = aQ - bQ^2$$

recall from earlier in the course, that this is a quadratic function (parabola). It is depicted in the following figure. The **slope of a quadratic function** has a formula which we discussed earlier (module 2). The slope of the total revenue curve is the **marginal revenue** curve. In this case, it is linear,

$$MR = a - 2bQ$$

and, we can see that the slope of the MR curve has the same intercept as the demand curve, and a slope that is twice as large. This is also depicted in the figure below.

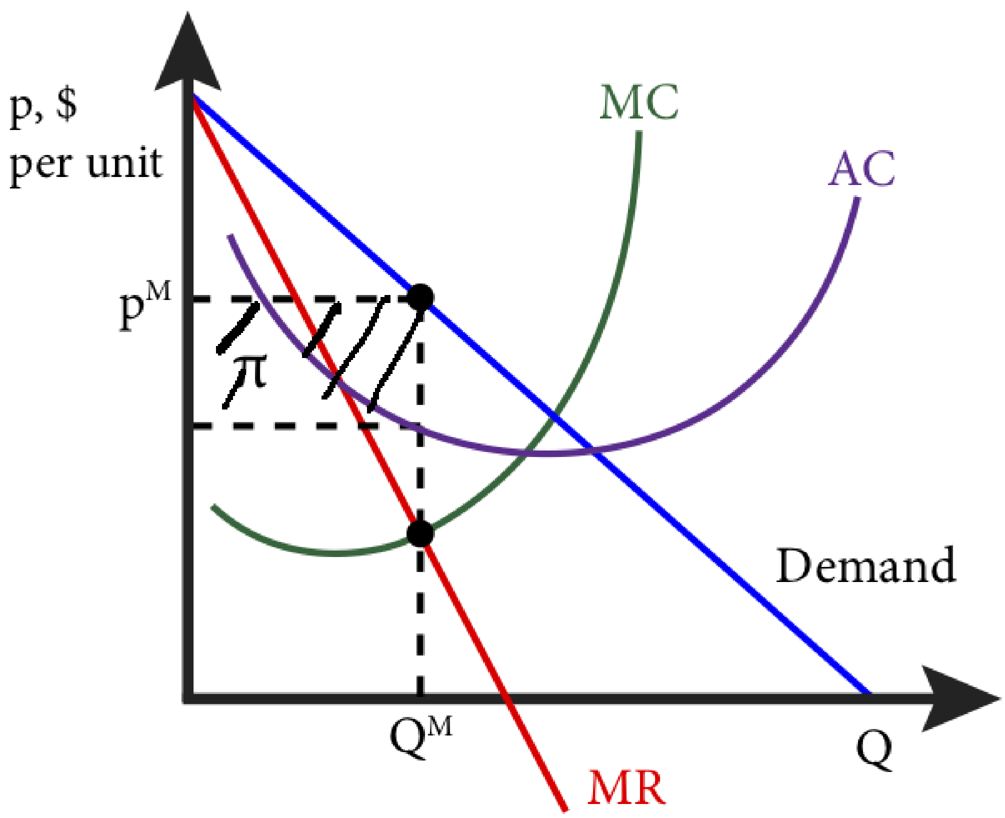
Figure: Marginal Revenue and Total Revenue for a Linear Demand Curve



Since marginal revenue is the rate of change in total revenue for a marginal increase in quantity, when marginal revenue is positive, the total revenue curve has a positive slope, and when marginal revenue is negative, the total revenue curve has a negative slope. Note that this means the total revenue is maximized at the point where marginal revenue is zero. This was how we derived the maximum of a quadratic function in Module 2 earlier in the course. We also did exercises about maximizing total revenue (for a nonprofit selling tickets to an event). That was a monopolist maximizing total revenue.

A profit maximizing firm, however, does not maximize total revenue. If there are any variable costs, maximizing total revenue implies too much production. Instead, a profit maximizing firm maximizes profits, which are total revenue *minus* total cost. To do so, it chooses output such that the marginal cost of producing the next unit of output exactly matches the marginal revenue it gets from selling the next unit of output. Now that we have the marginal revenue all that remains is to add the marginal cost. The figure below shows the optimal output for the monopolist where marginal revenue and marginal cost intersect. The monopolist shown as a marginal cost curve that has a curved U-shape; MC curves are often drawn this way in microeconomics courses.

Figure: Profit Maximization for a Monopolist



The intersection of the marginal revenue and marginal cost curves occurs at point Q^M and the highest price at which the monopolist can sell exactly Q^M is p^M. At Q^M the total revenue is simply $(p^M) \times (Q^M)$, and total cost is $(AC) \times (Q^M)$ since AC is just TC/Q. Therefore the shaded area is the difference between total revenue and total cost, that is, profit π.

Mathematically, profit maximization for a monopolist is the same as for any firm: find the output level for which marginal revenue equals marginal cost.[2]

Let's consider a specific example. Suppose that Tesla is a monopolist in the manufacture of luxury electric cars and faces an annual demand curve for its electric Model S cars of $Q = 200 - 2p$ in thousands, and p is in thousand \$s. Its fixed startup cost to producing Model S, which applies if it produces any of Model S but not if it chooses to produce none, is \$1 billion per year (think of this as a large lump sum payment converted to an annualized version for comparison with the other annual costs and revenues). Assuming it pays the fixed startup cost and produces some output, it has a marginal cost of producing a Model S of MC = Q, also in thousand \$s.

First, we will solve for the monopolist's profit maximizing level of output and the profit maximizing monopolist's price, assuming it produces a positive level of output. Second, we will graph the solution. Third, we will compare its total profits ignoring fixed costs to the fixed startup costs of \$1 billion. That comparison will determine whether it enters the market.

To solve the problem we need to proceed in four steps: one, find the inverse demand curve; two, find the marginal revenue curve; three, set marginal revenue equal to marginal cost and solve for Q^M, the monopolist's profit maximizing output level; four, find p^M the monopolist's profit maximizing price.

1) We find the inverse demand curve by solving for demand curve for p:

$Q = 200 - 2p$

$2p = 200 - Q$

$p = 100 - \frac{1}{2} Q$

[2] It is also important that marginal cost is increasing at the chosen quantity, otherwise the firm would be minimizing profit.

2) We find the marginal revenue curve by doubling the slope coefficient (B):

$MR = 100 - (2)\times(\frac{1}{2})\, Q$

$MR = 100 - Q$

3) Set MR equal to MC and solve for Q:

$MR = MC$

$100 - Q = Q$

$100 = 2Q$

$Q^M = 50$

or, converting to thousands,

$Q^M = 50{,}000$

4) Find p^M from the inverse demand curve:

$p = 100 - \frac{1}{2}\, Q$

$p = 100 - \frac{1}{2}\,(50)$

$p = 100 - 25$

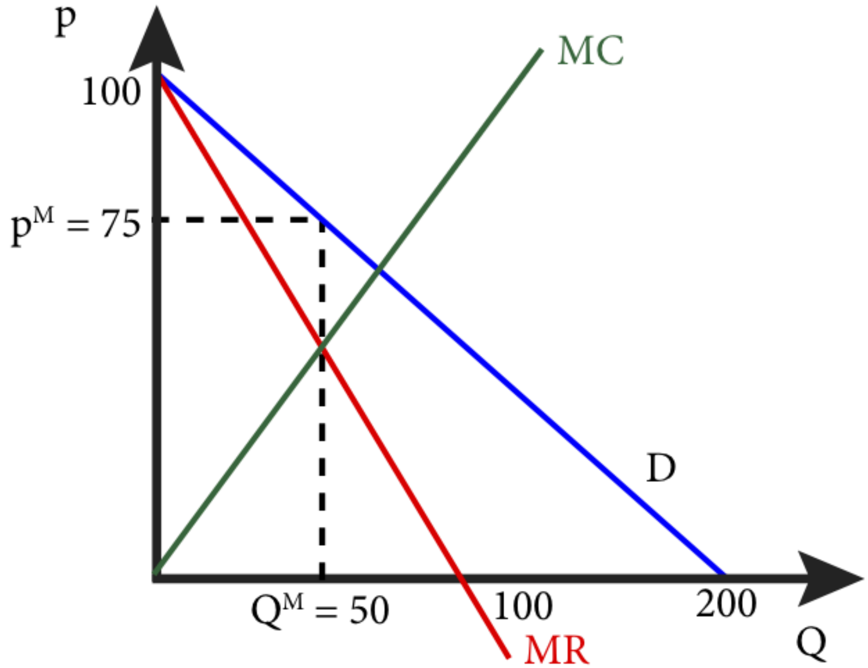
$p^M = \$75$

or, converting to thousands,

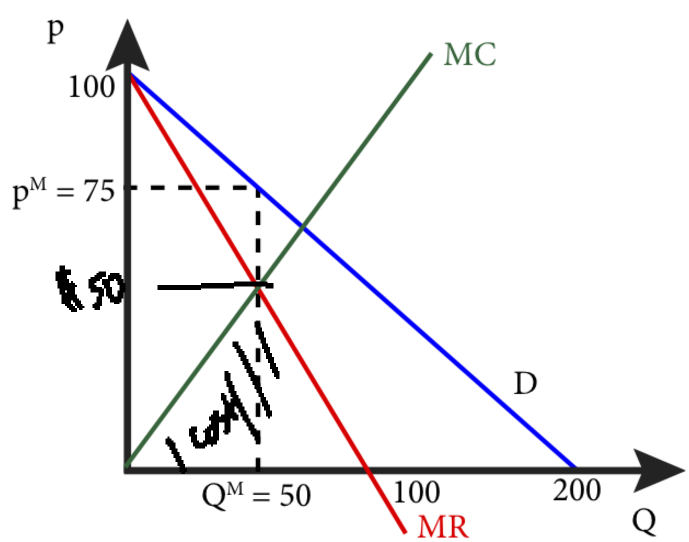
$p^M = \$75{,}000$

So the optimal output of Model S cars for Tesla is 50,000 cars and the optimal price is $75,000 per car. This is shown graphically in the figure below, where both axes are in thousands. This yields TR of $3.75 billion.

Figure: Solution to Tesla's Profit Maximization Problem



In order to determine Tesla's profit, we need to figure out total cost. The area of the shaded triangle in the picture below is the cost of production *not including the fixed startup cost*. That is $\frac{1}{2}50000\times50000\ =\ \$1.25$ billion. The fixed startup cost was $1 billion. The profit is therefore 3.75 – 1.25 – 1 = $1.5 billion. Thus, because profits are positive, Tesla enters this market.



In general, the total variable production costs are $\frac{1}{2}Q^2$ million dollars (the area of the triangle in the above figure), the total fixed production costs are 1000 million and the total revenue is $Q\times\left(100 - \frac{1}{2}Q\right)$ million, thus the profit function in millions is

$$\pi = Q\times\left(100 - \frac{1}{2}Q\right) - \frac{1}{2}Q^2 - 1000$$

Simplifying this expression on WolframAlpha yields

$$\pi = -Q^2 + 100Q - 1000$$

Which is a quadratic in Q. Optimizing a quadratic yields profit-maximizing quantity

$$Q^* = -\frac{b}{2a} = \frac{-100}{2*(-1)} = 50\ (thousand\ cars)$$

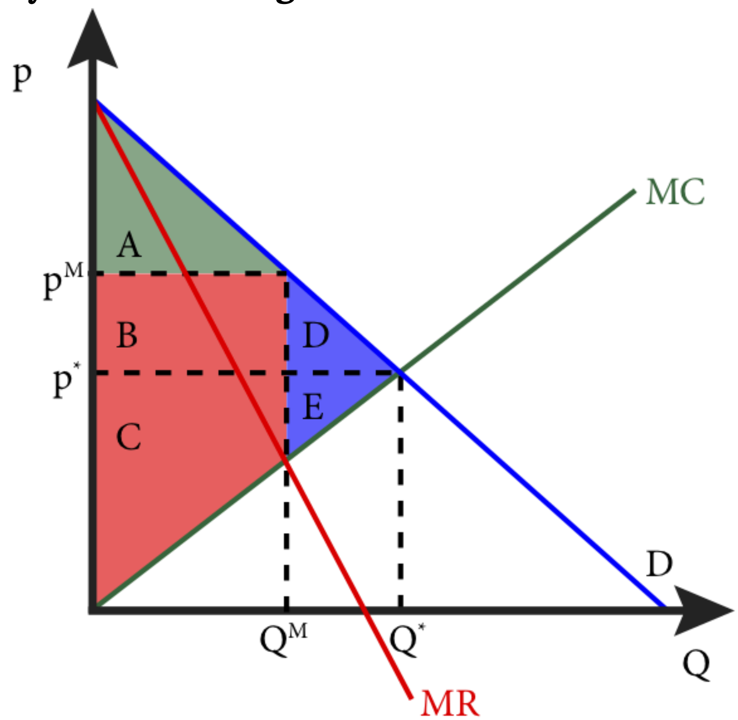
Which is what we found earlier by setting MR = MC. This can be plugged into the expression for π to get total profit, which is $1.5 billion.

Keep in mind that it possible in a particular case that we do all this math and yet, at the optimum, the monopolist makes a negative profit (i.e., a loss), because it cannot cover the fixed production cost at any single price. Then our prediction is that the monopolist will shut down (in the long run), or not enter the market. The product will not be produced. In the example we just went through, the firm did make a profit, and so we predict the product will be produced (at the monopoly quantity and price).

Implications of Monopoly for Efficiency and Equity

Earlier in the course, we used the sum of consumer and producer surplus as a measure of efficiency. Although this measure has criticisms, we will continue to use it in this section. Perfectly competitive markets that meet a set of criteria maximize surplus, and achieve efficiency (all mutually beneficial transactions take place). Monopoly markets do not. Intuitively, the monopolist's incentive is to limit output in order to keep prices high for all of its goods. But by limiting output, some surplus-increasing transactions do not take place (transactions where the buyer's willingness-to-pay exceeds the cost of production). Thus, monopoly leads to deadweight loss.

Figure: Monopoly and Deadweight Loss



In the Figure above, the monopolist output, Q^M, and the monopolist's price, p^M, create an area of consumer surplus of A. The producer surplus is area B+C and the total surplus is A+B+C. We can compare this outcome to the surplus-maximizing outcome which is shown as p^* and Q^*, or where price equals marginal cost. At this price, all transactions that have marginal cost less than the buyer's willingness-to-pay occur. Notice I do not say "willingness-to-accept" is less than "willingness-to-pay," because the monopolist's willingness-to-accept is higher than marginal cost: in order to be willing to produce another unit, the monopolist would have to be compensated for the additional cost of production *plus* the loss in revenue due to the necessity of lowering the price it obtains for the previous units. But the comparison for the calculation of surplus should be between the buyer's willingness-to-pay and the marginal cost of producing the unit, since (assuming social costs equal private costs, as we have been) the marginal costs of production reflect the loss to society from producing the unit. It is the market power of the monopolist that drives a wedge between the monopolist's willingness-to-accept and the cost of production.

I find it easier to think of p* as the surplus-maximizing price, rather than the price that would occur "if this monopolist were converted to a perfectly competitive market" (which is language sometimes used). It is very hard to imagine a breaking up of a monopoly into many small firms in such a way so that the sum of the supply curves of the many small firms will be equal to the marginal cost curve of this one firm. In general, the cost structures will not add up exactly. Still, taken the monopoly firm's cost structure and market structure as given, p* is the surplus-maximizing price, and it is not the price the monopoly picks.

At the surplus-maximizing price, the consumer surplus is A+B+D and the producer surplus is C+E for a total surplus of A+B+C+D+E. The difference in total surplus between the surplus-maximizing outcome and the monopoly outcome is D+E. This is the deadweight loss that results from the market power of the monopolist in this market. Again, the deadweight loss is the potential surplus not realized due to the lower level of output, which the monopolist picks in order to increase its profits.

In addition to a loss in efficiency, there is a potentially large equity concern. The monopolist takes area B from consumer surplus, leaving the consumers with less surplus. Monopoly can therefore be harmful for consumer welfare. If we, due to equity considerations, value monopoly profits less than consumer surplus, then this is a direct negative consequence of monopoly.

I encourage you to compare this monopoly picture with our earlier pictures of (the ideal) price floor or supply-side quota restriction. The monopoly problem is identical to a price control or supply-side quota restriction. The monopolist is picking a "price floor" (equivalently a quota) in order to maximize its total surplus. One important difference is that we can be confident that a profit-maximizing monopolist will pick the ideal case of these restrictions, since it bears all production costs. So we don't have to worry that we are understating the deadweight loss in this case. Still, when the government sets a price floor on a market to increase producer surplus (think of agricultural price controls), one interpretation is that the government is effectively stepping in to make the industry behave more like a monopolist.

As an example for calculating deadweight loss, suppose that a drug company has a patent for a drug that makes it a monopolist on the market for that drug. The demand for that drug is described by the inverse demand curve $p = 30 - Q$ and the firm's marginal cost is MC = Q. Where Q and p are in thousands. The monopolist's profit maximizing level of output is where MR equals MC and marginal revenue is a line with the same vertical intercept, 30, and twice the slope, 2. MR = 30 – 2Q.

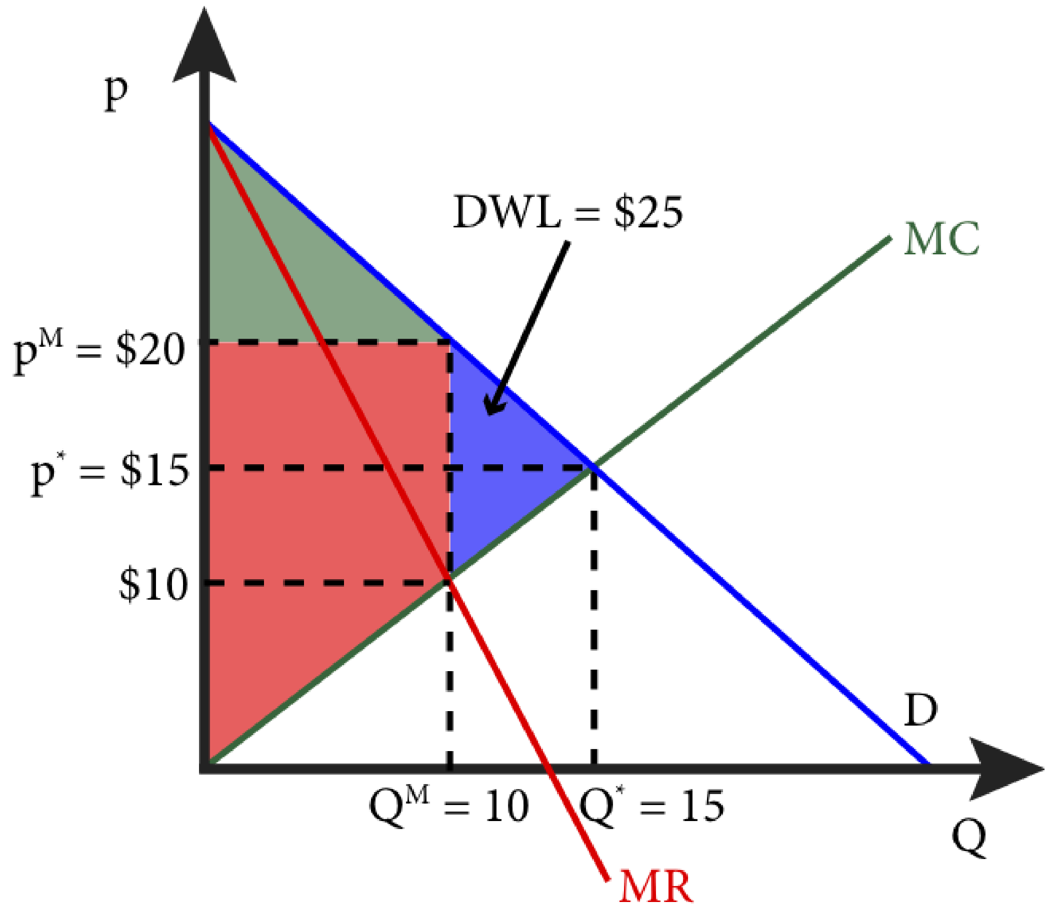
So MR = MC yields 30-2Q = Q, or 30 = 3Q, or $Q^M = 10$ million, and therefore, since $p = 30 - Q$, $p^M = \$20$.

That is the monopolist's profit maximizing solution. What is the efficient price? It is setting price equal to marginal cost, or

$p = 30 - Q = Q = MC$, or $2Q = 30$, or $Q^* = 15$ million and $p^* = \$15$

Graphically this looks like the following figure.

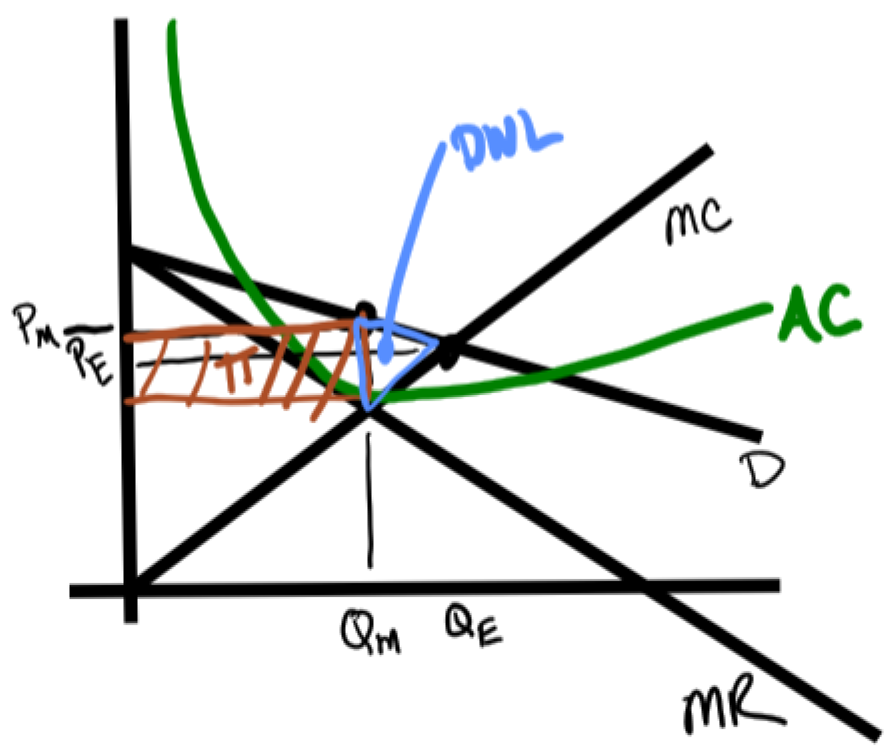
Figure: Deadweight Loss from Monopoly Example



We can see that the dead weight loss is a triangle with a base of 10, because at Q^M=10, MC=$10 and p^M=$20, therefore the distance between the two is $10, and a height of 5, the difference between Q^M and Q^*. Using the formula for the area of a triangle of ½ base times height, we get ½($10)(5) = $25 (million).

Market Power and the Demand Curve

A firm that faces a downward sloping demand curve, rather than a horizontal demand curve, has **market power**: the ability to choose a price above marginal cost. Monopolists face downward sloping demand curves because they are the only supplier of a particular good or service and the market demand curve is therefore the monopolist's demand curve. How much market power a firm has is a function of the shape of the demand curve. A market in which customers are very price sensitive is one with a highly elastic demand curve or one that is relatively flat. A market in which customers are very price insensitive is one with a highly inelastic demand curve or one that is relatively steep. A monopolist can get away with charging a higher price relative to marginal cost when the demand curve is steeper (i.e., price insensitive). Total monopoly profits, however, are also related to the size of the market. But for a given market size, we can expect monopoly profits to be higher when the demand curve faced by the monopolist is less price sensitive. The following plot shows a monopolist facing a relatively flat demand curve. As the demand curve gets flatter, the gap between the MR curve and the demand curve gets smaller at the profit-maximizing monopoly quantity. Thus, the difference between P_M and P_E gets smaller, and so does the DWL. For a horizontal demand curve (perfectly elastic), the monopoly price is P_E, and we are back to the perfectly competitive (price-taking) firm model. The degree of market power that the monopolist possesses can therefore be measured by how price-sensitive the consumers are that the monopolist faces: more price-sensitive implies less market power.



What makes customers more or less price sensitive? A number of factors, but the most important one is the availability of good substitutes. A drug company who has a patent for a drug that is the only cure for a serious disease is likely to face an inelastic demand curve as those that have the disease will be likely to buy at any price they can afford since there is no other choice. However, a cable company that has a natural monopoly in an area would probably face a highly elastic demand curve because customers would be likely to switch to other entertainment options if the price was too high.

Regulating Monopolies

Given the deadweight loss associated with monopolies – as well as possible inequities as consumer surplus is transferred to the monopolist – governments have an interest in regulating them. One common approach is to leave the monopolist alone. This works in many cases where the monopolist does not control a good that is deemed very important, or controls a good that has fairly close substitutes (so the monopolist doesn't have that much market power anyway); usually these two cases amount to the same thing.

Another common approach is to break up monopolies or prevent mergers and acquisitions. Ideally, when there are more firms in an industry, there will be greater competition and the result will be more like the competitive market ideal. If the monopoly exists despite no obvious cost advantage or other useful reasons (e.g. patents), or due only to very weak cost advantages, then this is probably a good strategy.

But suppose that breaking up the monopolist is not feasible, or not a good plan because breaking it up would lead to higher production costs. E.g., think about an electricity distributor – it would be inefficient to have two sets of electrical lines to every house. In this case, what can policymakers do?

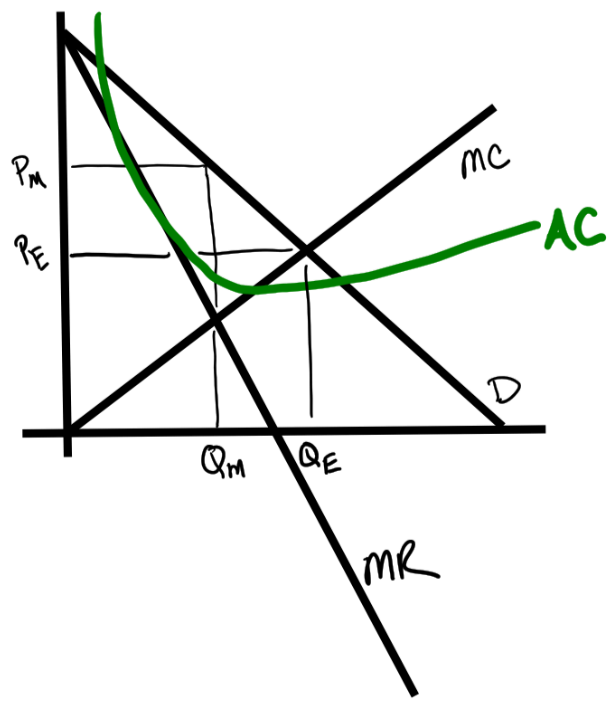
Traditionally there have been two popular alternative approaches. The regulated private enterprise approach, used relatively more often in the United States, is to let the firm be a monopoly, but impose a price ceiling on the firm. The public enterprise approach, used relatively more often in

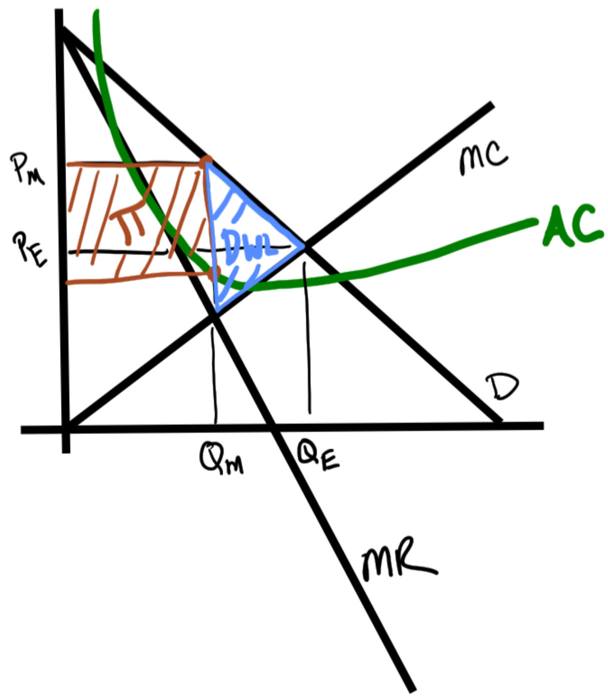
Europe, is to take over the monopoly and have the government run the enterprise. The public enterprise approach has been extensively compared with the regulated private enterprise approach, both theoretically and empirically (see e.g. Viscusi, Jr, and Sappington 2018) But only in a few cases do there appear to be arguably clear winners, so it seems hard to come up with a theory or good rule-of-thumbs about when to use one approach versus the other. The main source of conflict in the debate is over what the incentives are for the managers of the public enterprises to keep costs down and, in general, exert effort. The incentives for private enterprise on this front are much better understood. It is fairly well-accepted in these discussions that *if* public enterprise acted to maximize efficiency, then public control would indeed be efficient. But many political scientists tell me that assuming benevolent government is not realistic.

If policymakers cannot be trusted to run enterprises, then it is not immediately clear why they can be trusted to pick the right price ceilings in monopoly regulation. One argument in favor of this assumption is that imposing a price ceiling is probably an easier task than running a full enterprise.

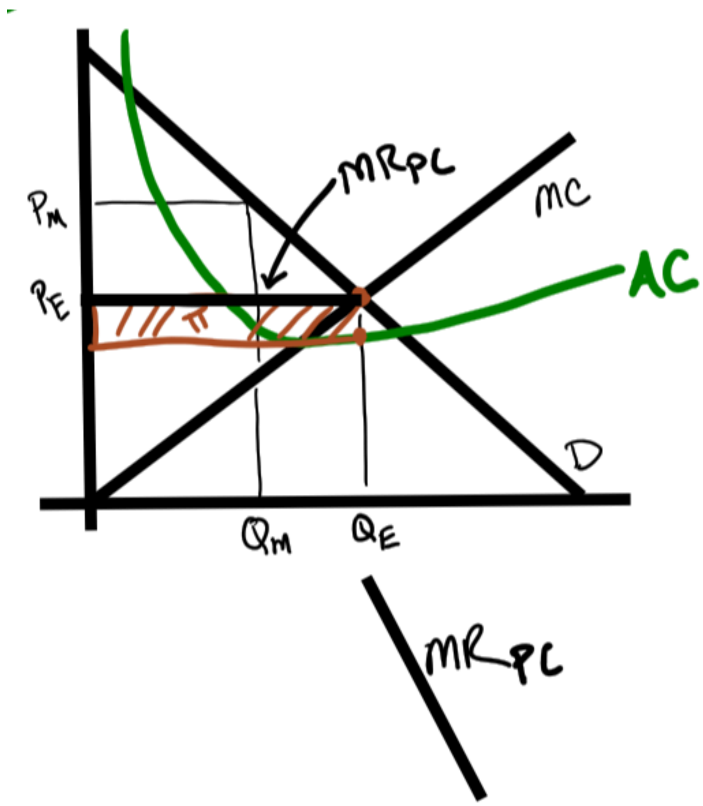
Price Control Regulation

In this module, we will focus on the price regulation approach, since we can study it using the monopoly model we have learned so far. To see that price control might be worthwhile, consider the following graph depicting a monopoly's cost and revenue curves, and the monopolist's decision to price at P_M. This monopoly pricing leads to deadweight loss as well as monopoly profits; these are shaded areas DWL and Π in the next figure. Notice that this monopolist has decreasing average cost for some range of production, due to economies of scale; the monopolist this picture represents has a large ("fixed") setup cost needed to produce the first unit, then increasing and linear marginal cost thereafter. If the government imposed a price ceiling on this monopolist, so that the monopolist could not charge above P_E – the efficient price – then the monopolist would charge exactly that. This is depicted in the third figure below. It can't charge higher (by assumption), and charging lower would only reduce its profits. Thus, the monopolist would charge P_E and produce Q_E; one way to see this is to draw the marginal revenue curve under the price control, which is flat at the price control through Q_E, then reverts back to the original MR curve (the new MR is labeled MR_{PC} in the figure). Deadweight loss would be avoided. Notice that the monopolist would still earn profits, albeit lower profits.





P_M
P_E
π
DWL
MC
AC
D
Q_M
Q_E
MR



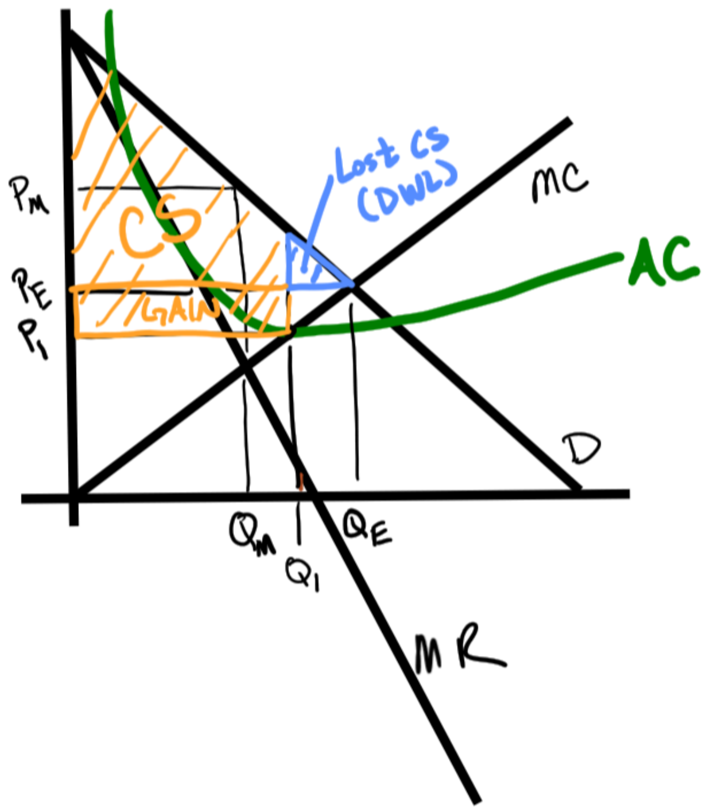
It is worth reflecting briefly on why the deadweight loss in the second of these figures is inefficient. If the monopolist produces Q_M, the monopolist already paying the fixed setup cost of production, that is needed to produce any positive quantity. At Q_M, an additional unit produced costs less than the willingness-to-pay of the consumers (the height of the demand curve). As a result, we say that it is efficient for this unit to be produced and sold, since a surplus over cost would be generated. However, it is not produced and sold because the monopolist does not want to cannibalize the revenue it makes on the people with higher willingness to pay. Consequently, it is not produced and sold. This is the source of the deadweight loss.

Note that even with the price ceiling of P_E, the monopolist earns a positive profit since the price exceeds the average cost of production at that point. If this were a story of competitive markets, we might predict market entry, which would dissipate those profits. But we are assuming in the monopoly story that there is no entry.

Tougher Price Controls for Equity?

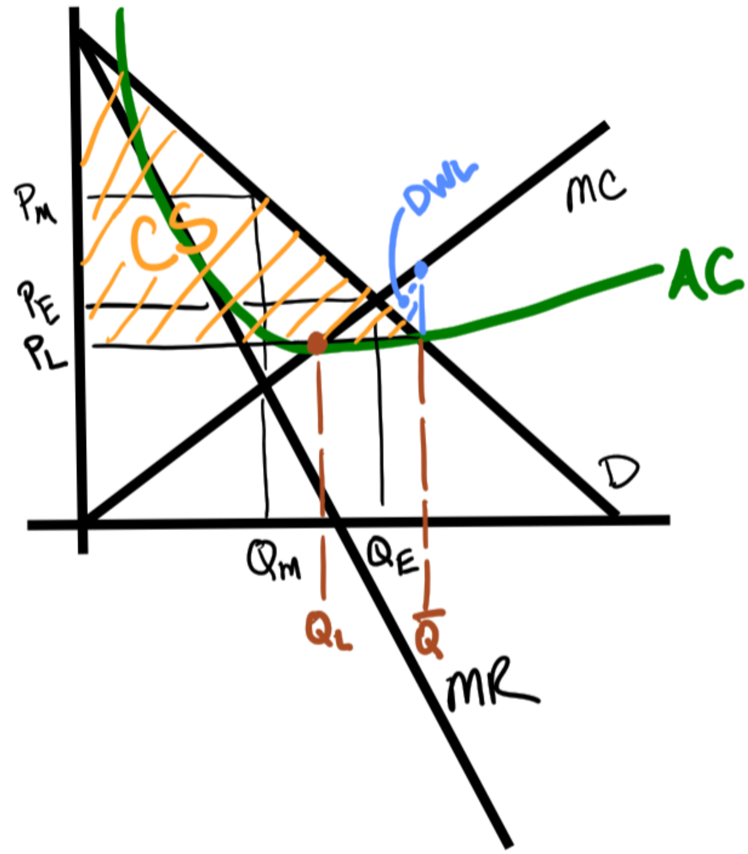
Can policy do better by pushing the price ceiling down further? It depends on the policy objective. If the objective is efficiency, then it cannot. The price P_E maximizes the total surplus, which counts a dollar in the hand of the monopolist equally to a dollar in the hand of a consumer. However, what if we instead wanted to maximize consumer surplus, ignoring the monopolist's surplus? This is an equity concern (based on my earlier definition of equity): we are now assuming a dollar in the hands of a consumer is valued at a dollar, while a dollar in the hands of the monopolist is valued at zero. Can we do better by pushing the price down?

With price regulation alone, the answer is "maybe." Consider the figure below. If policymakers use a price ceiling to push the price down to P_1, then the monopolist will respond by setting price P_1, and will produce where price (marginal revenue) equals marginal cost, which is quantity Q_1. The consumers lose the shaded triangle "Lost CS." The consumers gain the rectangle "GAIN." The rectangle is taken out of the monopolist's profits. If GAIN exceeds Lost CS, then this could increase consumer surplus. It might not, though. It depends whether the lost surplus to consumers shut out of the market due to the lower supply exceeds the gain to the consumers that remain in the market and can buy at the lower price-controlled price (we are essentially back to the old story of manipulating the surplus of consumers and producers via price controls).



If we also have the ability to do quantity controls, specifically requiring the monopolist to produce, then we can do even better for consumer surplus. Consider the following figure. This price ceiling is to push the price down to P_L, which is chosen to be where demand intersects the average cost curve (D intersects AC at point $\left(\overline{Q}, P_L\right)$). Left to its own devices, the monopolist would then produce at Q_L, where the forced price equals marginal cost. It is difficult to see in the picture, but the monopolist would make a sliver of profit by producing at Q_L, because average cost is lower than P_L there.

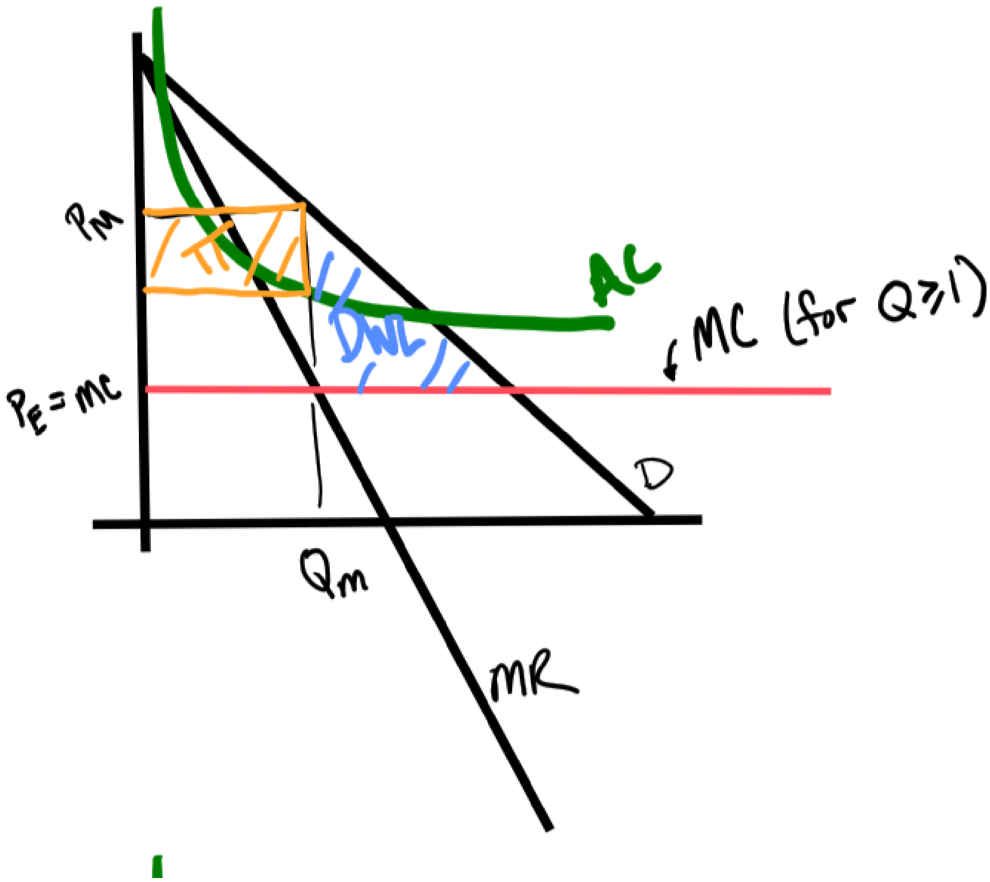
The next step is to impose a quantity control that takes away that last sliver of the monopolist's profit, and in the process expands consumer surplus. Suppose that policymakers mandate that the monopolist produce at least $\overline{Q}$. Then the monopolist produces $\overline{Q}$ and sells it at price P_L and just breaks even, because P_L was chosen to be the monopolist's average cost at that point. This twin controls policy maximizes consumer surplus; the consumer surplus triangle cannot be made any larger without putting the monopolist out of business (and losing all the CS).



With these policies we have managed to transfer all the monopoly profits to consumers, and get the maximum goods produced and sold in this market, but at what social cost? To do so, we have forced the monopolist to produce at $\overline{Q}$. The units produced between Q_E and $\overline{Q}$ are inefficient, in that the buyer's willingness to pay for those units is less than the marginal cost of producing them. Therefore, there is a deadweight loss from this control policy highlighted in the above figure as DWL. The deadweight loss occurs because some consumers will now buy the good despite the cost of producing the extra goods they buy being higher than the consumer's willingness-to-pay. If the private costs that the monopolist pays out are equal to social costs, this is a net social loss. But, this is not surprising, since these consumers still get some consumer surplus from the purchases at the controlled price, and our whole discussion here was about maximizing the consumer surplus generated by this market.

Regulating a Natural Monopoly without Bankrupting It

The previous example had the monopoly earning positive profits at the efficient price. However, that need not be the case. The following figure depicts a **natural monopoly**. This monopoly has average cost that always decreases in quantity. Marginal cost is constant beyond the first unit, so there must be some fixed cost for producing one unit. This could be, for example, a story of an electricity distributor; then the fixed cost, which is the marginal cost from going from 0 to 1 unit produced, is setting up the electrical grid, and the marginal cost thereafter (which is shown in the graph) is the cost of sending each unit of electricity through the lines, which we're assuming is constant to and through quantities easily encompassing the entire market. Then the most cost-efficient way to produce is to have one firm produce at maximum scale, and serve the whole market. The figure also depicts the monopoly price and quantity, monopoly profits (which are positive), and deadweight loss.



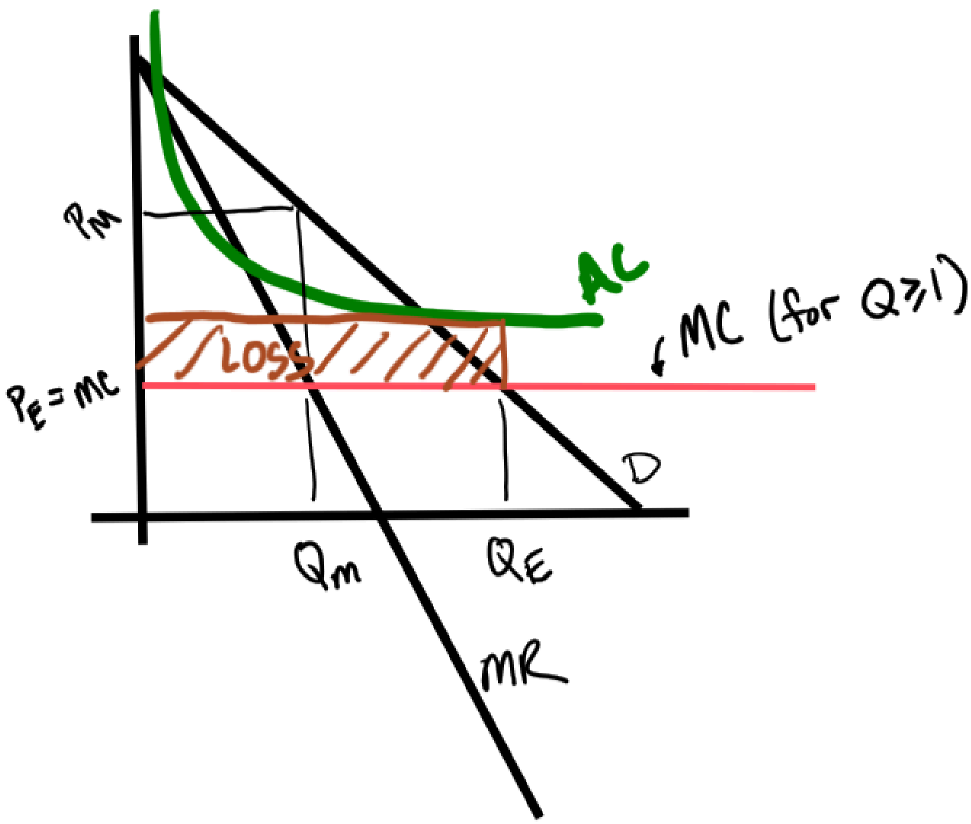
It is first worthwhile to discuss why competition "should" be socially wasteful in this particular case. If a competitor entered with the same technology, then the competitor would have to pay the fixed cost, but would produce at the same marginal cost. This means that the resources for the fixed cost payment are spent without any gain. It's like paying for a redundant set of electric poles that are not needed. Now, it is possible that a competitor would enter anyway – that depends on whether or not, if the two firms split the demand, they could both earn nonnegative economic profits. But then we'll be talking about a duopoly; those stories are left to the next module. In any case, this competition would be "inefficient," in the sense that if we could just get this one firm to "behave," there should be no need for a redundant set of electric lines!

Without government intervention however, there's no reason for this firm to behave. As a monopolist, it would charge the monopoly price P_M so that the quantity produced Q_M is where MC = MR, as shown in the figure above. The deadweight loss is shown as DWL, and monopoly profits as π.

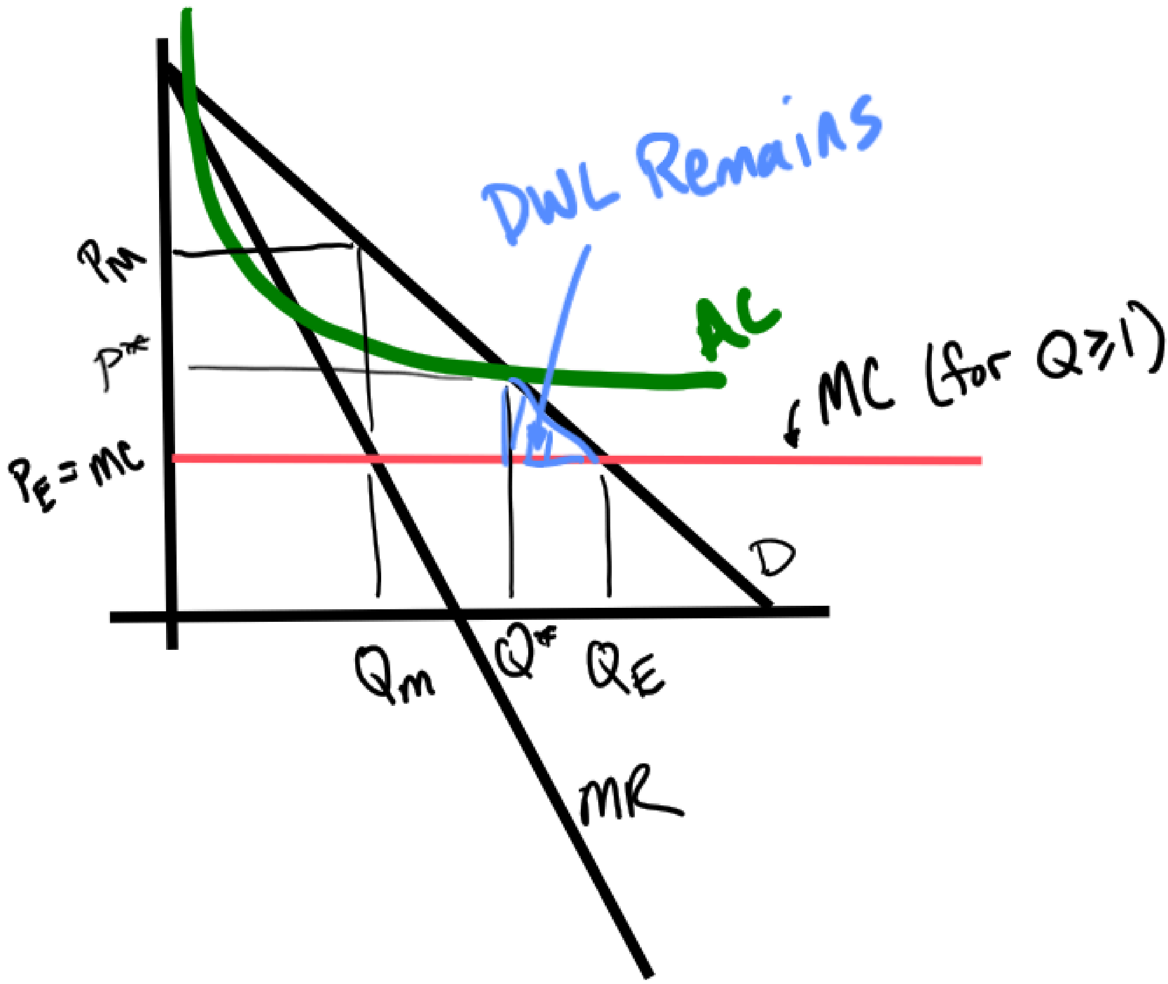
The government might consider instituting a price ceiling. Let's suppose the government sets price at exactly marginal cost, in an attempt to achieve efficiency. But then, the firm would operate at a loss! This is depicted below. Setting price equal to marginal cost does not permit the monopolist to cover the fixed cost, and so the monopolist would shut down or not enter in the first place (in the long-run).

This highlights a regulatory issue in the case of natural monopoly: to maximize consumer surplus, the regulator may be tempted to force the monopolist to price at marginal cost. But by doing so that leaves the monopolist no surplus over variable cost to recover the fixed costs. We can also think of this as a commitment problem. Once the private power company has built the power plant, surplus-maximizing government regulators may be tempted to force the price below what the private power company needs to recover fixed costs. The power company will be worse off, but if the power plant investment is sunk, them being worse would not change behavior, and therefore would not lead to inefficiency. In the long-run though, such pricing would deter future power plant entrants and have a chilling effect on investment in these natural monopolies.[3]

[3] This kind of commitment problem on the part of government could also be used to evaluate nationalization of private industries. In the short-run, nationalizing private industry could benefit society at the expense of the private firms holding those investments. In the long-run, private investment will be deterred by a country's history of nationalization.



This suggests that it may be tricky to use price ceilings to regulate a natural monopolist. The concern is that the government will choose a price ceiling that is too low for the natural monopolist to recoup the upfront, fixed investment in infrastructure. If the price ceiling is set too low, the monopolist won't be able to cover its fixed cost, and will not enter the market (or will eventually leave). The best the government can do in this case is set the price ceiling to P^*, shown in the following picture.



This is the price that sets average cost equal to demand. At this price, the monopolist makes no profits – it is just able to recoup the fixed cost and the necessary variable costs for producing at this amount. However, this price still leaves some deadweight loss, as shown in the picture.

It is useful to discuss why this area is deadweight loss. The consumers in that range of the demand curve have a willingness to pay that is below P^*, so they don't buy the good. But the cost of servicing them is MC, which is below their willingness to pay. Consequently, these are transactions that would be socially beneficial if they *could* take place – the demanders have willingness-to-pay in excess of the cost to make them the product. But these transactions, in this story, cannot take place, because if the government tried to force the price low enough for them to take place, then (by the assumption of one price) *all* of the previous purchasers would be able to pay that lower price. Because of the great loss in revenue due to that latter force, the monopolist wouldn't be able to recoup the fixed costs, and would go out of business.

This raises the question: could there be a way to pick these relatively-low-but-still-high-enough willingness-to-pay consumers out of the population of consumers, and then charge them a special rate? If so, then surplus gains could be made from that. But to do this, one would have to be clever enough to avoid spoiling the rest of the market. Being that clever is called *price discrimination*, that is, posting different prices for different customers. Price discrimination takes us beyond the simple monopoly model, which assumed only one price chosen, the same for all customers, and we do not cover price discrimination in this course.

References

Friedman, Lee S. 2002. *The Microeconomics of Public Policy Analysis*. Princeton, N.J: Princeton University Press.

Viscusi, W. Kip, Joseph E. Harrington Jr, and David E. M. Sappington. 2018. *Economics of Regulation and Antitrust, Fifth Edition*. fifth edition. Cambridge, Massachusetts; London, England: The MIT Press.

License

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

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