

Notes 2: Theory of the Firm

I. The Neoclassical Theory of the Firm

A neoclassical firm is an organization that **controls** the transformation of **inputs** (resources it owns or purchases) into **outputs** (valued products that it sells), and **earns** the difference between what it receives in revenue and what it spends on inputs.

A. Production Technologies: Inputs and Outputs

1. Technology sets

The technology set for a given production process is defined as

$$T = \{(x, y) : x \in \mathbb{R}_+^n, y \in \mathbb{R}_+^m : x \text{ can produce } y\}$$

where x is a vector of inputs and y is a vector of outputs. The set consists of those combinations of x and y such that y can be produced from the given x . As an example consider the production technology for producing pancakes on a weekend camp-out. The input vector might be as follows:

$$x = \begin{pmatrix} \text{powdered milk} & \text{water} & \text{eggs} & \text{oil} & \text{flour} \\ \text{baking powder} & \text{salt} & \text{bowl} & \text{whip} & \text{measuring devices} \\ \text{small griddle} & \text{camp stove} & \text{white gas} & \text{spatula} & \text{semi-skilled labor} \\ \text{butter} & \text{maple syrup} & \text{plate} & \text{knife} & \text{fork} \end{pmatrix}$$

The output in this case is a single product consisting of buttered pancakes covered with syrup ready to eat. The technology set then consists of different numbers of pancakes along with all the various input combinations that could produce them. For later reference denote this as technology 1. One element of this set might be as follows:

$$\begin{pmatrix} 1/3 \text{ c powdered milk} & 15/16 \text{ c water} & 1 \text{ egg} & 2 \text{ T oil} & 1 \text{ c flour} \\ 2 \text{ t baking powder} & 1/4 \text{ t salt} & 1 \text{ bowl} & 1 \text{ whip} & 1 \text{ measuring set} \\ 1 \text{ small griddle} & 1 \text{ camp stove} & 1/4 \text{ c white gas} & 1 \text{ spatula} & 1/4 \text{ h semi-skilled labor} \\ 3 \text{ T butter} & 1/2 \text{ c maple syrup} & 1 \text{ plate} & 1 \text{ knife} & 1 \text{ fork} \end{pmatrix} \quad (10 \text{ pancakes})$$

Of course, the same inputs with an output of 6 pancakes is also possible since we can always throw the extras to the wild creatures (assuming we are not particularly environmentally conscious). Other combinations are also possible. A particular element of the production set is called a **production plan**.

The production process for pancakes can of course be defined in different ways depending on which parts we want to consider. If the output is pancakes hot off the griddle, then the inputs butter, maple syrup, plate, knife and fork can be eliminated. We also might consider dividing the process up into steps where the first step is the production of "pancake mix". In this case the technology for hot off the griddle pancakes might be

$$\left[\begin{array}{cccc} \text{pancake mix} & \text{water} & \text{butter} & \\ & \text{bowl} & \text{whip} & \text{measuring set} \\ \text{small griddle} & \text{camp stove} & \text{white gas} & \text{spatula} \end{array} \right] \text{ (pancakes) } \left[\begin{array}{c} \\ \\ \text{semi\&skilled labor} \end{array} \right]$$

where the pancakes are assumed to be off the grill only. This might be denoted technology 2. We could also consider a more primitive process denoted technology 3 that does not use the manufactured input flour but considers wheat, a grindstone and grinding labor as additional inputs replacing flour.

2. Returns from the production technology

The returns to a particular production plan are given by the revenue obtained from the plan minus the costs of the inputs or

$$p \cdot \sum_{j=1}^m p_j y_j - \sum_{i=1}^n w_i x_i$$

where p_j is the price of the j 'th output and w_i is the price of the i th input.

A neoclassical firm is then an organization that controls one or more production technologies and receives the returns from implementing those technologies in a particular way, paying for the inputs, and receiving the revenue from the outputs.

3. Objectives of a firm

We typically assume that a firm exists to make money. Such firms are called *for-profit* firms. Given this assumption we can set up the firm level decision problem as maximizing the returns from the technologies controlled by the firm taking into account the demand for final consumption goods, opportunities for buying and selling products from other firms and the actions of other firms in the markets in which the firm participates. In perfectly competitive markets this means the firm will take prices as given and choose the levels of inputs and outputs that maximize profits. In the case where the firm controls a single technology within a vertical chain the problem can be written

$$\max_{x,y} \left[p y - \sum_{i=1}^n w_i x_i \right] \text{ such that } [(x,y) \in T]$$

If the firm controls more than one production technology it would take into account the interactions between the technologies and the overall profits from the group of technologies. A firm may choose to acquire more technologies and control more steps in the chain if that will lead to lower costs of producing and marketing the product within the chain.

4. Relationship of a firm to a vertical chain

As discussed previously, a **vertical chain** is the process that begins with the acquisition of raw materials and ends with the distribution and sale of finished goods. Within a vertical chain a firm may control only one production technology such as the transformation of a particular set of raw materials into an intermediate product or it may control several of the technologies. Within a particular vertical chain it is common to consider only technologies that have a single output. If the technology inherently has **more than one output** we say the technology is **joint**. In such cases we have to consider the impacts of this jointness on costs and returns. Such cases will be discussed in more detail later.

The firm may choose to organize the technologies that it controls in a variety of ways. Consider again the example of the pancakes. The firm may choose to use technology 1 or technology 2. In the case of technology 2 the firm could produce its own pancake mix or it could purchase it on the market. The **vertical boundaries** of a firm in a vertical chain define the activities that a firm performs for itself as opposed to purchasing them from independent firms in the market. Activities closer to the beginning in a vertical chain are called upstream in the chain while those closer to the finished goods are called downstream. Thus a firm's vertical boundaries deal with how many stages up or downstream from a given process the firm chooses to control.

5. Expendable inputs and capital inputs

The factors of production used by a firm fall into two general classes, those that are used up in the production process and those that simply contribute a service to the process. For example the flour that goes into pancakes is gone once the pancakes are made and sold, while the mixing bowl is still available for future use. Thus we categorize inputs into two categories, expendables and capital.

- a. **Expendable factors of production** are raw materials, or produced factors that are completely used up or consumed during a single production period. Common examples of these factors that lose their identity with a single use are seed, fuel, lubrication, some pesticides and fertilizer, feed, and feeder animals.
- b. **Capital** is a stock that is not used up during a single production period, provides services over time, and retains a unique identity. Examples include machinery, buildings, equipment, land, stocks of natural resources, production rights, and human capital. Since the capital is not used up we define the service provided by the capital as the input that is used.
- c. **Capital services** are the flow of productive services that can be obtained from a given capital stock during a production period. They arise from a specific item of capital rather than from a production process. It is usually possible to separate the right to use services from ownership of the capital good. For example, one may hire the services of a potato harvester to dig potatoes, a laborer (with embodied human capital) to provide milking services for a given period, or land to grow crops.

A number of examples will illustrate the argument. Land is considered a capital asset but the right to use the land for a specific period is an expendable service flow. A laborer and the embodied human capital is considered capital, but the service available from that laborer is considered an expendable capital service. Shares in an irrigation company are considered capital but the acre feet available for use in a given season are an expendable input.

B. Output Sets and Efficient Use of Inputs

1. Definition

Rather than representing a firm's technology with the technology set T , it is often convenient to define the output set for a given technology. The output set $P(x)$ is the set of all output vectors $y \in R_+^m$ that are obtainable from the input vector $x \in R_+^n$. Specifically we say that

$$P(x) = \{y : (x, y) \in T\}$$

If there is only one output, then $\max P(x)$ is the maximum level of y that can be produced using a given level of x . The firm figures out how to "optimally" use the level of resources x and no more output can be obtained by combining them in another way. Each input is being used in such a way it cannot produce more output.

2. The efficient subset of $P(x)$

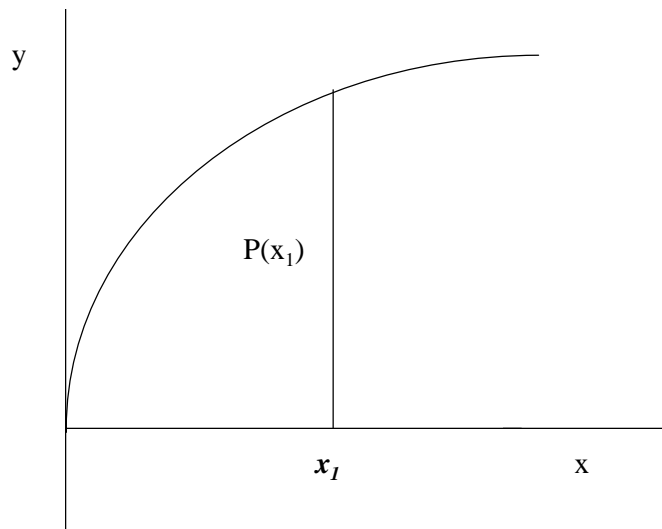
The efficient output subset is defined as follows:

$$\text{Eff } P(x) = \{y : y \in P(x), \text{ and } \nexists y' \in P(x) \text{ such that } y' \succ y\}$$

In essence, the efficient set is elements of $P(x)$ such that any expansion in any element in the output y will remove it from $P(x)$. See part 4 below for an illustration.

If there is only one output, then $\text{Eff } P(x) = \{\max P(x)\}$

3. Graphical representation with one input and one output

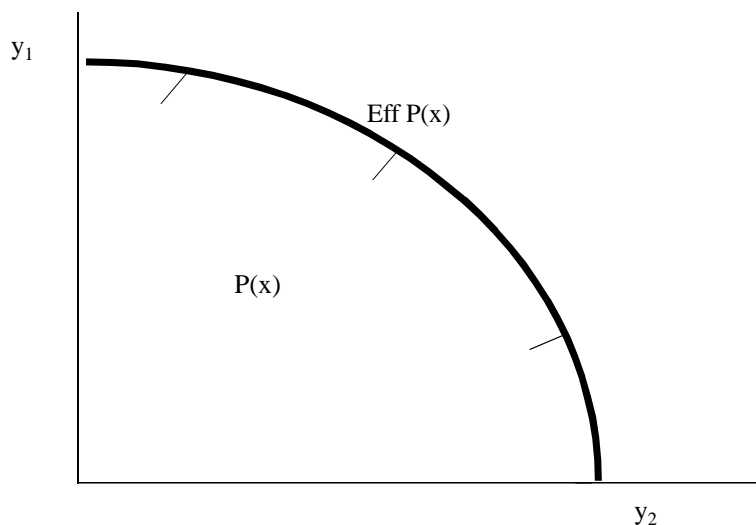


In this case, $P(x)$ is a line segment for each x . Production is efficient only along the curve which lies above all of these line segments. Points below the curve represent less output with the same level of input.

4. Graphical representation with two outputs

In the case of two outputs the output set is not just a line segment, but a region of the plane, the set of all combinations of y_1 and y_2 that can be produced with given levels of the x variables. As an example consider the various combinations of corn and corn silage that can be produced on a given acre of land with a fixed set of inputs.

Again, production is efficient only along the upper boundary of $P(x)$, or $\text{Eff } P(x)$, defined above.



5. Optimal use of inputs

- a. A firm uses engineering, agronomic, accounting, economic and other principles in order to insure that it is on the boundary of the output set.
- b. It turns out that if a firm maximizes profit for a given set of prices, it will necessarily be on the boundary of the output set.
- c. The optimal organization of inputs is sometimes called “technical efficiency.”

C. Cost minimization

In a world of certainty (where the firm's production function is not random) the maximization of profit implies the minimization of cost. Thus a firm can consider the profit maximization process in two steps. In the first step the firm decides the least cost way to produce various levels of output. In the second step the firm chooses the optimum output level.

1. We can define the cost minimization problem for the firm as follows:

$$C(y,w) = \min_x \sum_{j=1}^n w_j x_j$$

such that $(x,y) \in T$

This gives the minimum cost at which the output level y can be produced, given the technology T and input prices w .

We define $x(y,w)$ to be the input combination at which this minimum cost is obtained.

2. Examples

- a. Hog rations

Consider a farmer who is mixing hog rations for his finishing unit. A typical input vector might be as follows:

$$\left(\begin{array}{cccccc} \text{tractor} & \text{grinder\&mixer} & \text{fuel} & \text{corn} & \text{milo} & \\ \text{soybean meal} & \text{mineral premix} & \text{vitamin premix} & \text{antibiotic supplement} & \text{semi\&skilled labor} & \end{array} \right)$$

To produce $y = 2000$ pounds of a finished ration with certain nutritional attributes, one choice of inputs might be

$$\left(\begin{array}{cccc} 1/4 \text{ hr tractor time} & 1/4 \text{ hr grinder\&mixer time} & 1 \text{ gallon fuel} & 1720 \text{ lbs corn} \\ 0 \text{ lbs milo} & 224 \text{ lbs soybean meal} & 46 \text{ lbs mineral premix} & 5 \text{ lbs vitamin premix} \\ 5 \text{ lbs antibiotic supplement} & 1/4 \text{ hr semi\&skilled labor} & & \end{array} \right)$$

The cost minimization problem is then to choose the combination of these inputs that will produce the 2000 pounds of complete ration at least cost. This choice will depend on the prices of the inputs and their nutritional content. For any y , there will be many input vectors x such that (x,y) is in T .

- b. Apple storage

Consider a fruit packer considering possible ways to store apples. The output is pounds of apples stored for a certain number of days with a minimum level of spoilage. The inputs include alternative types of storage equipment including simple cooled rooms and more complex climate controlled systems, labor, fuel, boxes, etc. Depending on the quality of the apples, the time to be stored and the prices of the inputs, the packer will choose alternative methods of storage.

3. Representing the optimum graphically

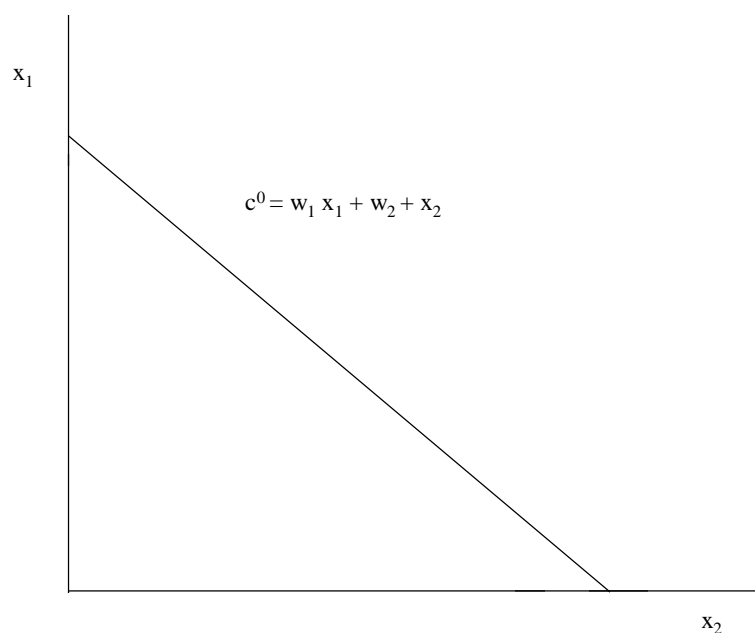
a. isocost lines

The cost of using a given set of inputs is given by the equation

$$\text{Cost} = \sum_{j=1}^n w_j x_j$$

$$= w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

If we set the cost equal to a constant we get the equation for a line if there are two inputs, or a plane if there are three inputs, etc. Graphically this can be represented as



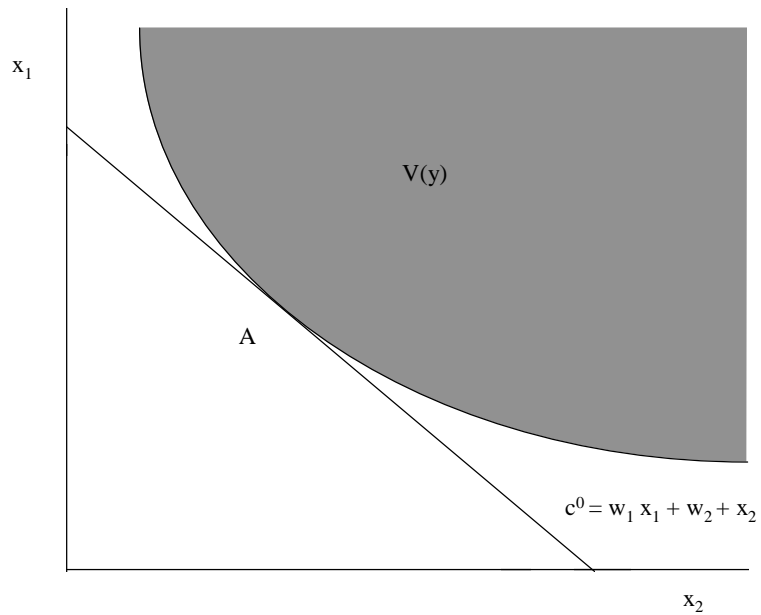
Points to the northeast of the line have higher cost than c^0 while points to the southwest have lower cost. With a different set of prices, w , the line will have a different slope.

b. Finding the minimum cost of a given output level

For any y , there are many possible input combinations which can be used. The task is to find the combination of inputs that has the lowest cost.

Define $V(y) = \{x \text{ such that } (-x, y) \in T\}$

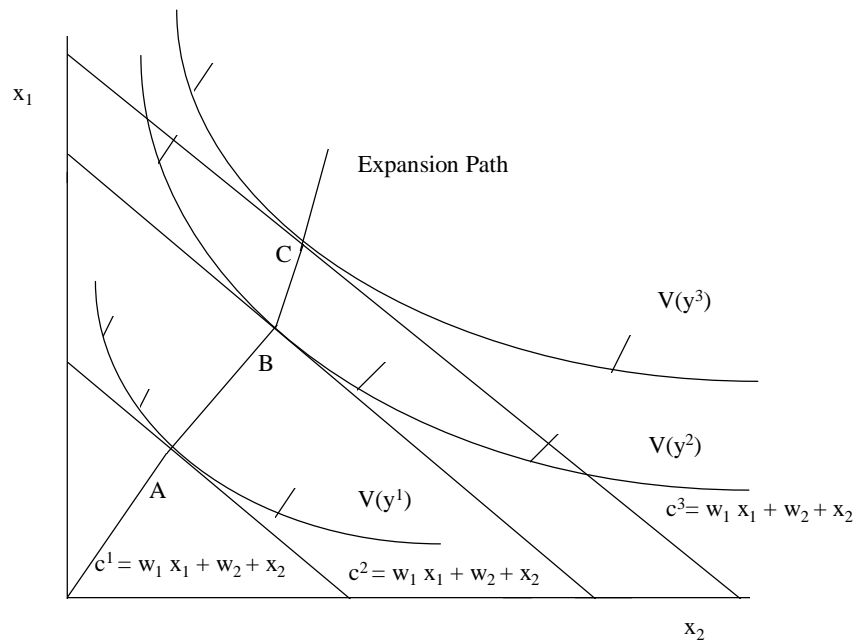
This is the set of all input combinations which may be used to produce y . The cost of any point in the figure (x_1, x_2) is given by $w_1 x_1 + w_2 x_2$. The minimum cost of producing y can be obtained graphically by finding the point at which $V(y)$ is tangent to an isocost line. This minimum is obtained at $A = x(y, w)$, where it is clear that any points in $V(y)$ other than A lie to the northeast of the isocost line c^0 .



As an example consider the various combinations of corn, corn silage, soybean meal, milo, hay, molasses, and a mineral supplement that can be used to produce 5 tons of cattle feed with a specific protein and net energy content.

4. The expansion path of the firm

For any given output level and set of input prices, there is a least cost combination of inputs $x(y,w)$. As output changes, there will be different combinations of inputs that minimize cost. Starting from one level of y , the set of cost minimizing input combinations as output expands, is called the expansion path of the firm. Graphically the expansion path might look as follows:



The firm initially minimizes cost at point A with a cost of c^1 . When required to produce the amount y^2 the firm changes input levels to point B and incurs cost c^2 . The firm then moves to point C when the output expands to y^3 .

D. Analyzing the Costs of Production

1. Opportunity cost

Above, we assume that an input price vector w exists. But how do we determine the costs of these inputs? Since the firm often must choose whether to buy inputs or make them itself, the concept of opportunity cost is important. **The opportunity cost of any good or service is its value in its next best alternative use.** For example, the opportunity cost of the service of an input used in the production of any particular commodity is the maximum amount that the input would produce of any other commodity.

In order to make quantities of different goods comparable, we measure opportunity costs in terms of a common unit, usually money. Thus, the opportunity cost of any good or service is the maximum amount the good or service could receive elsewhere, for use as a production input or for final consumption. When a market transaction is not available to value a given expendable input or the service of a capital input, we attempt to approximate the opportunity cost. Usually, we form these approximations by determining implicit rental rates.

For example, when active cash rental markets for land exist, these rental rates provide a good estimate of the cost of land use. When cash rental arrangements are not common, share rental rates can be used to approximate the actual factor cost. An active market in general purpose buildings does not normally exist. Similarly, active markets may exist for unskilled workers, so that we can use commonly reported wage rates; but the market for skilled managers may be much smaller, and thus we must use of opportunity cost calculations in this case.

If a producer sells some of an expendable produced input to his neighbor for a documented price and also uses some of the same expendable factor in his own production of another output, then the value of this output may be used to value the factor. Consider, for example, a dairy farmer who produces more corn silage than he needs for his dairy herd, and who sells the excess to his neighbor, who picks it up on the farm. The price the neighbor pays may be used to measure the value of corn silage to the dairy enterprise on the same farm.

In the case of capital inputs that last more than one period, an implicit rental rate must often be determined if there is no regular rental market. The most common way to do this is to assume that markets are efficient and that purchase prices represent the discounted service use values.

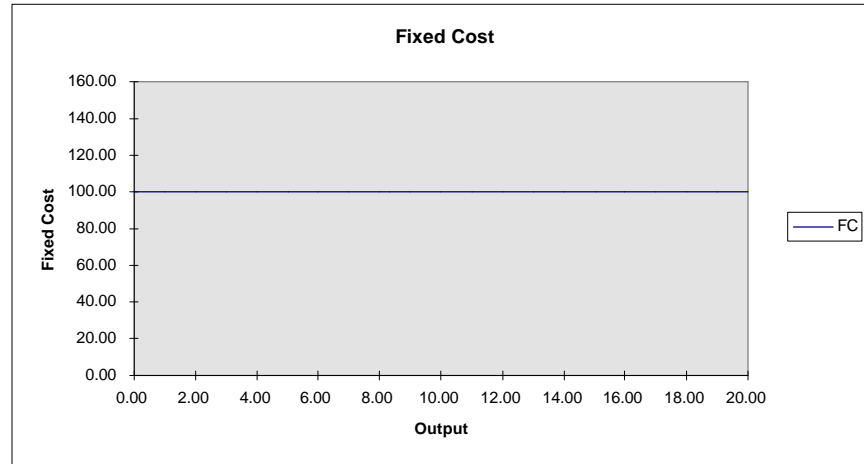
2. Fixed cost

Fixed costs are those costs that the firm is committed to pay for factors of production, regardless of the firm's current decisions. If some choices are fixed for a given decision problem, then costs associated with them are also fixed. Sometimes we say that fixed costs are those costs that do not vary with the level of output. Examples of fixed costs are business licences, annual lease payments on land, ownership costs such as time dependent depreciation and implicit interest on machinery and equipment, and fertilizer already applied to a growing corn crop. We often refer to fixed costs with the abbreviation FC.

Consider the following cost function for a firm:

$$c(y) = 100 + 6y + .4y^2 + .02y^3$$

Fixed costs for this cost function are 100. Graphically fixed costs can be represented by a horizontal straight line.



The fixed costs, FC, are equal to 100, independent of the level of y chosen.

3. Sunk and avoidable cost

In some situations it is possible for the firm to recoup part of its fixed costs if it decides to go out of business. For example, a farmer with a fixed lease may be able to sub-lease the acreage to another producer and thus not lose the entire annual payment. Or a swine producer may be able to sell his farrowing stalls and some other equipment at the time of ceasing operations. But a firm may not be able to get a refund for its business license, and a farmer who has built a specialized grain handling system for his poultry operation may find no ready buyer.

We can subdivide fixed costs into two components:

$$\text{Fixed Cost} = \text{Sunk Cost} + \text{Avoidable Fixed Costs}$$

The portion of fixed costs that is not recoverable is called **sunk cost**. A sunk cost is like spilt milk: once it is sunk, there is no use worrying about it, and it should not affect subsequent decisions.

Costs, including fixed costs, that are not incurred if operations cease, are called **avoidable costs**. The portion of the cost of a grain truck that can be recouped by selling it at an auction is an avoidable cost.

4. Variable cost

Variable costs are those costs that are affected by the firm's actions in the current period. Variable costs occur because of the decision to purchase additional factors or factor services for use in production. Variable costs are usually abbreviated VC. Common examples of variable costs are fuel, labor, feed, and fertilizer. Variable costs are always defined in reference to a specific output level. Thus the variable cost of producing 50 tractors differs from the cost of producing 200 tractors. In all cases it is assumed that the firm has minimized the cost of producing this output level. Thus $VC(y,w)$

$= C(y,w)$ if all inputs are variable. As the output level changes, the optimal levels of the variable inputs and thus the cost will change. The **expansion path** of the firm traces out the effect of output changes on input use and the associated cost.

The time period under consideration clearly affects the delineation of fixed and variable factors and associated costs. For example, if a tractor is leased (with no possibility of releasing) on an annual basis, the cost of the lease is fixed (and sunk) when deciding whether to produce cotton or tomatoes, but the per acre charge for custom harvesting is variable when deciding whether to harvest a damaged crop. Once the owner of a resource decides to assume ownership for another period, the ownership costs, service reduction costs due to time, and potential market value gains are fixed. As irreversible decisions on input use are made, costs that were previously variable become fixed. In this vein, the cost of all expendable inputs are variable until they are contracted on for use. For operator-owned capital goods, costs are fixed once the operator decides to maintain the asset for another period.

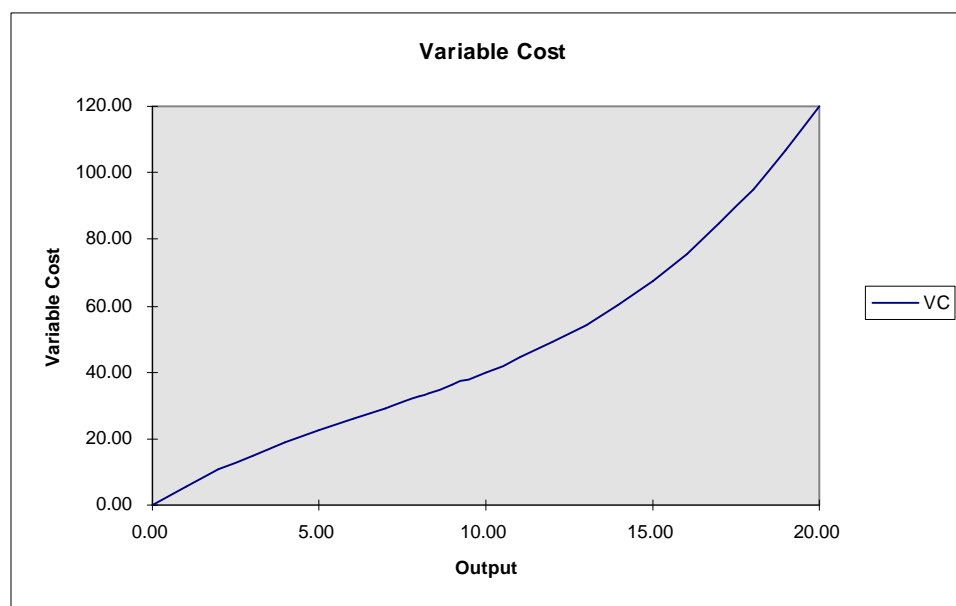
For the cost function

$$c(y) = 100 + 6y + .4y^2 + .02y^3$$

variable costs (VC) are given by

$$VC(y) = 6y + .4y^2 + .02y^3$$

Graphically variable costs vary as output varies. This curve comes from matching up the costs associated with an optimal set of inputs and output levels in an expansion path diagram.

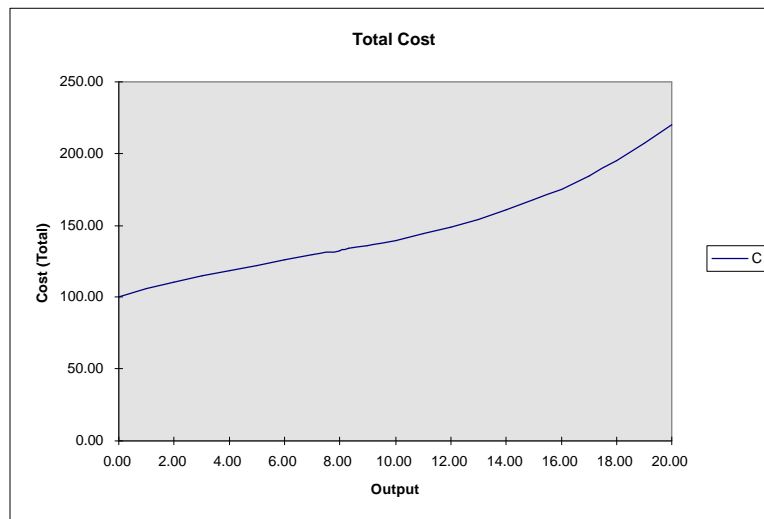


5. Total cost

The sum of fixed and variable costs is called **total cost** (C). Algebraically $C = FC + VC$. In all cases, it is assumed that the firm has minimized costs for this output level given its constraints. When some inputs are fixed, the cost function can be written $C(y,w,z) = FC(z) + C(y,w)$. Only the variable inputs

are included in C.

Graphically, total cost is just the vertical displacement of variable cost.



6. Marginal cost

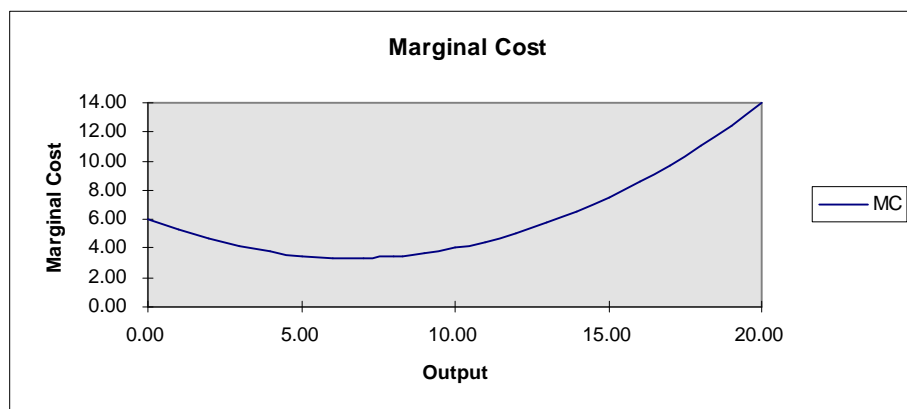
Marginal cost is the increment, or addition, to cost that results from producing one more unit of output. The increase in total cost is the increase in variable cost. Marginal cost is just the derivative of the cost function with respect to output. In discrete terms

$$MC = \frac{\Delta C(y,w)}{\Delta y}$$

while in infinitesimal terms

$$MC = \frac{dC(y,w)}{dy}$$

Graphically, marginal cost represents the slope of the cost curve.



a. Example with corn

Consider the marginal cost of producing corn on a per acre basis. The marginal cost represents the increment in cost from the additional inputs required to increase the expected yield by one bushel.

b. Numerical example

For the cost function

$$C(y) = 100 + 6y + .4y^2 + .02y^3$$

marginal cost is given by

$$\frac{dC(y)}{dy} = 6 + .8y + .06y^2$$

7. Average cost

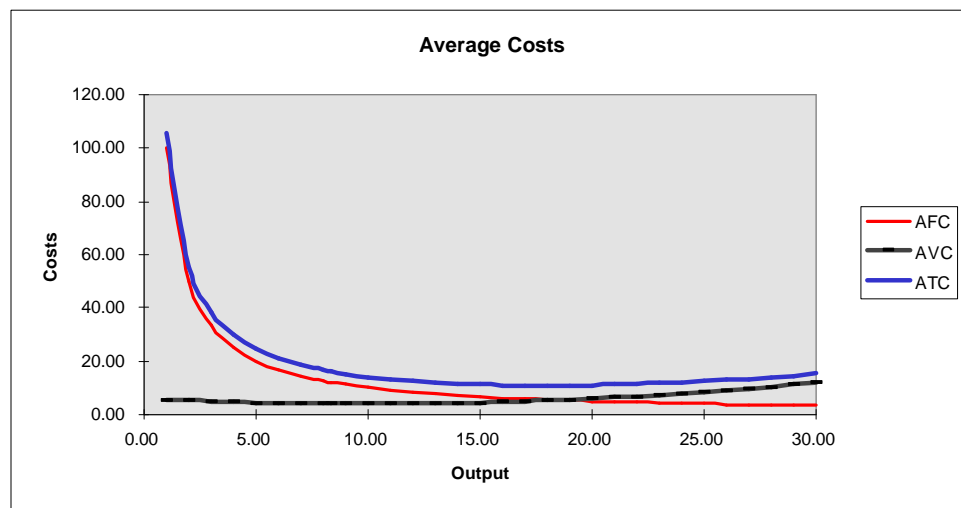
Average cost is just a total cost figure divided by the associated output level. Thus it represents the average per unit cost of that level of output.

Average fixed cost is given by $AFC = FC/y$

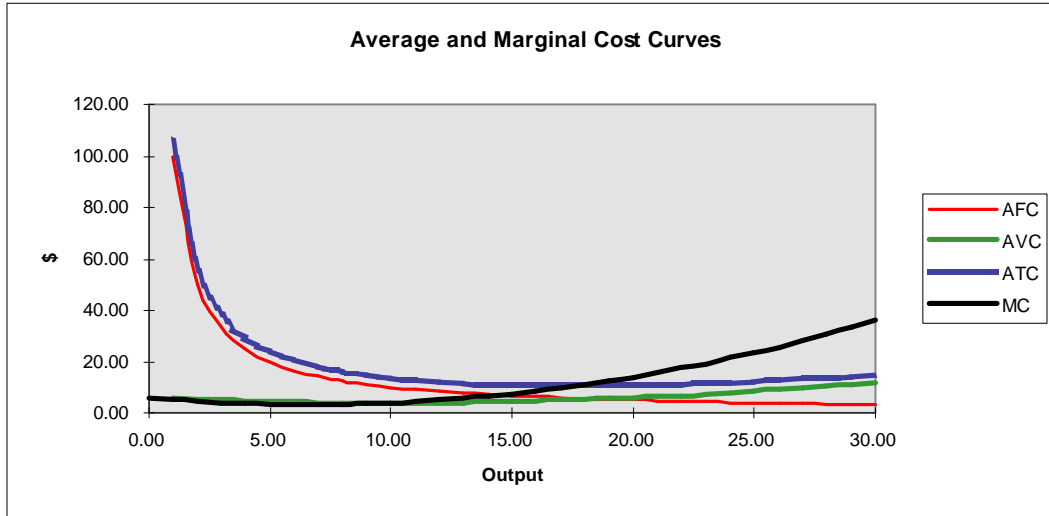
Average variable cost is given by $AVC = VC(y,w)/y$

Average (total) cost is given by $AC = C(y,w)/y$

The following graph shows these three types of average cost:



Since the marginal cost curve is the slope of the cost curve, it will be equal to average cost at points where average cost is flat. In the case of a U-shaped cost curve this means marginal cost will intersect average cost at its minimum as in the diagram below.



8. Consider the following table which is the basis for the above graphs.

Y	FC	VC	C	AFC	AVC	ATC	DMC	MC
0.00	100.00	0.00	100.00					6.00
1.00	100.00	5.62	105.62	100.00	5.62	105.62	5.62	5.26
2.00	100.00	10.56	110.56	50.00	5.28	55.28	4.94	4.64
3.00	100.00	14.94	114.94	33.33	4.98	38.31	4.38	4.14
4.00	100.00	18.88	118.88	25.00	4.72	29.72	3.94	3.76
5.00	100.00	22.50	122.50	20.00	4.50	24.50	3.62	3.50
6.00	100.00	25.92	125.92	16.67	4.32	20.99	3.42	3.36
7.00	100.00	29.26	129.26	14.29	4.18	18.47	3.34	3.34
8.00	100.00	32.64	132.64	12.50	4.08	16.58	3.38	3.44
9.00	100.00	36.18	136.18	11.11	4.02	15.13	3.58	3.66
10.00	100.00	40.00	140.00	10.00	4.00	14.00	3.82	4.00
11.00	100.00	44.22	144.22	9.09	4.02	13.11	4.22	4.46
12.00	100.00	48.96	148.96	8.33	4.08	12.41	4.74	5.04
13.00	100.00	54.3	154.34	7.69	4.18	11.87	5.38	5.74
14.00	100.00	60.48	160.48	7.14	4.32	11.46	6.14	6.56
15.00	100.00	67.50	167.50	6.67	4.50	11.17	7.02	7.50
16.00	100.00	75.5	175.52	6.25	4.72	10.97	8.02	8.56
17.00	100.00	84.66	184.66	5.88	4.98	10.86	9.14	9.74
18.00	100.00	95.04	195.04	5.56	5.28	10.84	10.38	11.04
19.00	100.00	106.8	206.78	5.26	5.62	10.88	11.74	12.46
20.00	100.00	120.0	220.00	5.00	6.00	11.00	13.22	14.00
21.00	100.00	134.8	234.82	4.76	6.42	11.18	14.82	15.66
22.00	100.00	151.4	251.36	4.55	6.88	11.43	16.54	17.44
23.00	100.00	169.7	269.74	4.35	7.38	11.73	18.38	19.34
24.00	100.00	190.1	290.08	4.17	7.92	12.09	20.34	21.36
25.00	100.00	212.5	312.50	4.00	8.50	12.50	22.42	23.50
26.00	100.00	237.1	337.12	3.85	9.12	12.97	24.62	25.76
27.00	100.00	264.1	364.06	3.70	9.78	13.48	26.94	28.14
28.00	100.00	293.4	393.44	3.57	10.48	14.05	29.38	30.64
29.00	100.00	325.4	425.38	3.45	11.22	14.67	31.94	33.26
30.00	100.00	360.0	460.00	3.33	12.00	15.33	34.62	36.00

9. Other factors in the cost function

When defining output in the cost function, we usually do not specify the time required to complete the process. In many instances the cost of producing an item will rise as the time required is reduced. For example, producing automobiles at a faster pace may put more stress on the workers, which will lead to more errors and thus to more rejects and a higher cost per unit of quality merchandise.

Production at a constant rate may be also cheaper than producing the same number of units, some at a faster rate and some at a slower rate. For example, the yield from combining corn may be higher per unit of combine time if production is uniform, rather than very fast some of the time and very slow

some of the time.

10. Short and long run costs

It is often useful to distinguish between costs in the short-run and costs in the long-run. What we mean by the “short-run” and the “long-run” depends on what costs we consider to be fixed. As time periods become shorter, more and more costs become fixed.

The **short-run** is a time period brief enough that the firm can vary some, but not all, of its inputs in a costless manner. If some inputs are fixed or cannot be varied in a costless manner, then the firm will minimize short-run costs.

The **long-run** is a time period long enough that the firm can vary all of its inputs in a costless manner. By costless we do not mean that inputs have zero price. Rather, we mean that there is no adjustment cost of varying the levels of these inputs. For example, installed feeding equipment is costly to remove and replace. Thus it is effectively a fixed input for short periods of time. When the firm has no fixed inputs and no adjustment costs, the firm will minimize long run costs.

If there are costs associated with varying the level of an input we say that the firm experiences **adjustment costs**. The most common type are associated with machinery and equipment. But there may be adjustment costs to laying off and rehiring workers as well.

II. Production Functions and Productivity

A. Representing technology with a production function

1. Representations of the production functions

- a. The **technology set** for a given production process is defined as

$$T = \{ (x, y) : x \in R^n, y \in R^m : x \text{ can produce } y \}$$

where x is a vector of inputs and y is a vector of outputs. The set consists of those combinations of x and y such that y can be produced from the given x . We assume that if $(x, y) \in T$, then for any $x' \leq x$, (x', y) is also in T . When input levels rise, you can always throw away inputs you don't need and produce the same level of output you produced before. This assumption is called **free disposal** of inputs. We also assume that if $(x, y) \in T$, then for any $y' \leq y$, (x, y') is also in T . If the input vector x is sufficient to produce the output vector y , then you can always throw away outputs you don't need and produce less output than before. This assumption is called **free disposal** of outputs.

- b. The **output set** $P(x)$ for a given technology is the set of all output vectors $y \in R^m$ that are obtainable from the input vector $x \in R^n$. Specifically we say that

$$P(x) = \{ y : (x, y) \in T \}$$

If there is only one output, then $\max P(x)$ is the maximum level of y that can be produced using a given level of x . The firm figures out how to “optimally” use the level of resources x and no

more output can be obtained by combining them in another way. Each input is being used in such a way it cannot produce more output.

- c. The input requirement set $V(y)$ of a given technology is defined as

$$V(y) = \{x: (x, y) \in T\}$$

This is the set of all combinations of the various inputs that will produce at least the level of output y . As an example consider the various combinations of corn, corn silage, soybean meal, milo, hay, molasses, and a mineral supplement that can be used to produce 5 tons of cattle feed with a specific protein and net energy content.

2. Definition of a production function

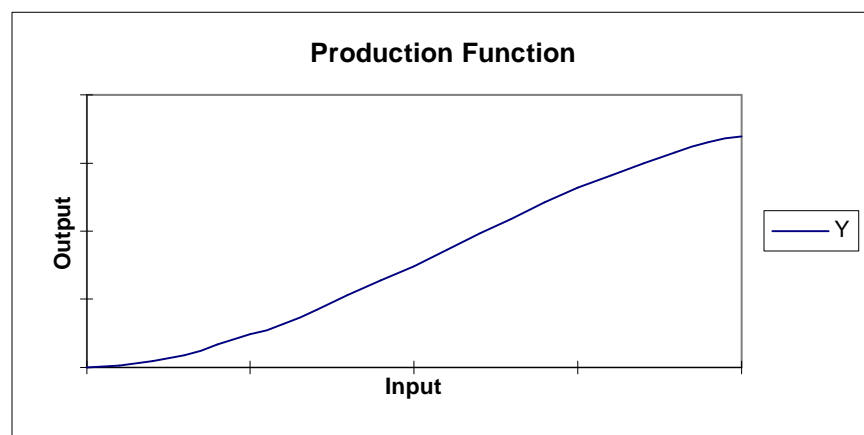
To this point we have described the firm's technology in terms of a technology set, the input requirement set $V(y)$ or the output set $P(x)$. For many purposes it is useful to represent the relationship between inputs and outputs using a mathematical function that maps vectors of inputs into a single measure of output. In the case where there is a single output it is sometimes useful to represent the technology of the firm with a mathematical function that gives the maximum output attainable from a given vector of inputs. This function is called a production function and is defined as

$$\begin{aligned} f(x) &= \max_y [y: (x, y) \in T] \\ &= \max_y [y: x \in V(y)] \\ &= \max_{y \in P(x)} [y] \end{aligned}$$

Once the optimization is carried out we have a numerically valued function of the form

$$y = f(x_1, x_2, \dots, x_n)$$

Graphically we can represent the production function in two dimensions as follows



In the case where there is one output, one can also think of the production function as the boundary of $P(x)$, i.e., $f(x) = \text{Eff } P(x)$.

3. Examples of production functions

- a. Consider the production technology for corn on a per acre basis. The inputs might include 1 acre of land and various amounts of other inputs such as tillage operations made up of tractor and implement use, labor, seed, herbicides, pesticides, fertilizer, harvesting operations made up of different combinations of equipment use, etc. If all but the fertilizer are held fixed, we can consider a graph of the production relationship between fertilizer and corn yield. In this case the production function might be written as

$$y = f(\text{land, tillage, labor, seed, fertilizer, } \beta)$$

- b. Numerical examples

Consider a production function with two inputs given by $y = f(x_1, x_2)$. It might have the form (also called the Cobb-Douglas form)

$$y = Ax_1^{a_1} x_2^{a_2}$$

$$= 5x_1^{\frac{1}{3}} x_2^{\frac{1}{4}}$$

or

$$y = a_1 x + a_2 x^2 + a_3 x^3$$

$$= 10x + 20x^2 + 0.60x^3$$

We usually postulate a simple functional form for the production function such as a polynomial or a power function.

4. Marginal product

The firm is often interested in the effect of additional inputs on the level of output. For example, the field supervisor of an irrigated crop may want to know how much crop yield will rise with an additional application of water during a particular period. For small changes in input levels this output response is measured by the marginal product of the input in question. In discrete terms the marginal product of the i th input is given as

$$MP_i = \frac{\Delta y}{\Delta x_i} = \frac{y^2 - y^1}{x_i^2 - x_i^1}$$

where y^2 and x^2 are the level of output and input after the change in the input level and y^1 and x^1 are the levels before the change in input use. For small changes in x_i the marginal physical product is given by the partial derivative of $f(x)$ with respect to x_i , i.e.,

$$MP_i = \frac{\partial f(x)}{\partial x_i} = \frac{\Delta y}{\Delta x_i}$$

This is the incremental change in $f(x)$ as x_i is changed holding all other inputs levels fixed.

Values of the discrete marginal product for the above production function are in the table. For example the marginal product in going from 4 units of input to 5 units is given by

$$MP_i = \frac{\Delta y}{\Delta x_i} = \frac{475 - 321.6}{5 - 4} = 153.40$$

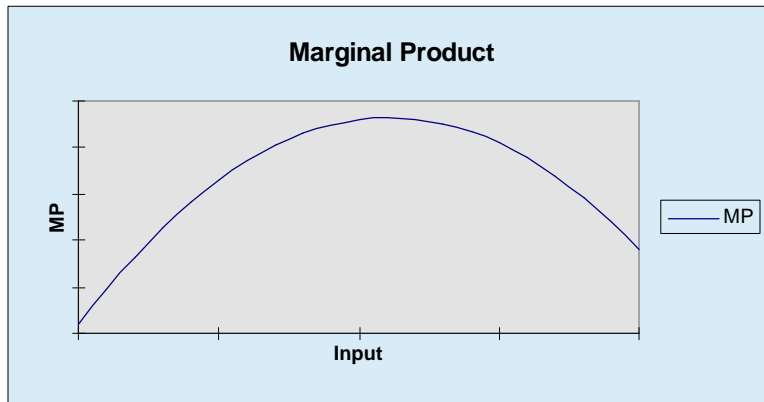
For the production function given by $y = 10x + 20x^2 + 0.60x^3$ we can compute the marginal product using the derivative as follows

$$\frac{dy}{dx} = 10 + 40x + 1.80x^2$$

At $x = 4$ this gives 141.2 while at $x = 5$ this gives 165.0 .

5. Graphical representation of the marginal product

The marginal product function for the production function pictured above is as follows.



Notice that it rises at first and then falls as the production function's rate of increase falls.

6. Average product

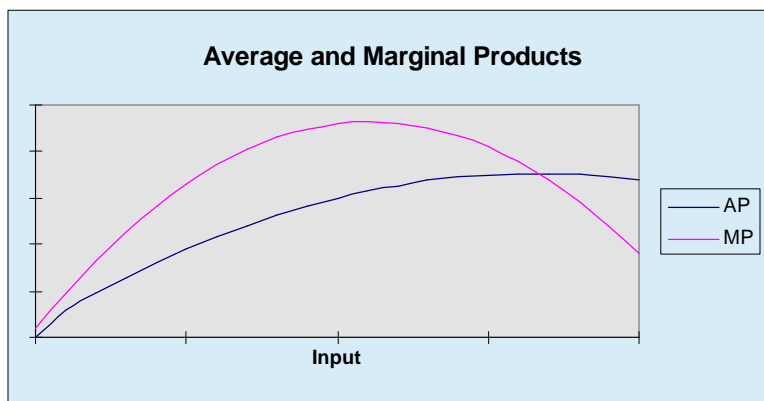
The marginal product measures productivity of the i th input at a given point on the production function. An average measure of the relationship between outputs and inputs is given by the average product which is just the level of output divided by the level of one of the inputs. Specifically the average product of the i th input is

$$AP_i = \frac{f(x)}{x_i} = \frac{y}{x_i}$$

For the above production function the average product at $x=5$ is $475/5 = 95$.

7. Graphical representation of the average product

For the production function given above, the average product function along with the marginal product function are as follows below. Notice that the marginal product curve is above the average product curve when the average product curve is rising. The two curves intersect where the average product reaches its maximum.



8. Tabular representation of production function

Tabular representation of production data for the function $y = 10x + 20x^2 + 0.60x^3$

Input (x)	Output (y)	Average Physical Product y/x	Discrete Marginal Physical Product $\frac{\Delta y}{\Delta x}$	Marginal Physical Product $\frac{dy}{dx} = 10 + 40x + 1.8x^2$
0.00	0.00			10.00
1.00	29.40	29.40	29.40	48.20
2.00	95.20	47.60	65.80	82.80
3.00	193.80	64.60	98.60	113.80
4.00	321.60	80.40	127.80	141.20
5.00	475.00	95.00	153.40	165.00
6.00	650.40	108.40	175.40	185.20
7.00	844.20	120.60	193.80	201.80
8.00	1052.80	131.60	208.60	214.80
9.00	1272.60	141.40	219.80	224.20
10.0	1500.00	150.00	227.4	230.0
11.0	1731.40	157.40	231.4	232.2
12.0	1963.20	163.60	231.8	230.8
13.0	2191.80	168.60	228.6	225.8
14.0	2413.60	172.40	221.8	217.2
15.0	2625.00	175.00	211.4	205.0
16.0	2822.40	176.40	197.4	189.2
17.0	3002.20	176.60	179.8	169.8
18.0	3160.80	175.60	158.6	146.8
19.0	3294.60	173.40	133.8	120.2
20.0	3400.00	170.00	105.4	90.0

B. Profit Maximization

1. General Problem

Most firms have an objective to maximize profits.

The firm-level maximization problem can be written as follows:

$$\max_{x, y} [py - \sum_{j=1}^n w_j x_j]$$

such that $(x, y) \in T$

where T is a description of the technology, i.e.,

$$T = \{(x, y) : x \in \mathbb{R}_+^n, y \in \mathbb{R}_+ : x \text{ can produce } y\}$$

The set T represents the combinations of inputs and outputs that are mutually compatible. The problem can also be written as

$$\begin{aligned} \max_{x, y} [py - \sum_{j=1}^n w_j x_j] \text{ such that } x \in V(y) \\ \text{or} \\ \max_{x, y} [py - \sum_{j=1}^n w_j x_j] \text{ such that } y \in P(x) \end{aligned}$$

where the technology is represented by $V(y)$, the input requirement set, or $P(x)$, the output set.

2. Profit maximization with a single output and a production function representation of technology

For the case of one output, the technology can be represented by a production function, and thus the maximization problem can be written

$$\max_x [pf(x_1, x_2, \dots, x_n) - \sum_{j=1}^n w_j x_j]$$

If the production function is continuous and differentiable we can use calculus to obtain a set of conditions describing optimal input choice. If we differentiate the above expression with respect to each input we obtain

$$p \frac{\partial f(x)}{\partial x_j} - w_j = 0, \quad j = 1, 2, \dots, n$$

Since the partial derivative of f with respect to x_j is the marginal product of x_j this can be interpreted as

$$p \text{MVP}_j = w_j, \quad j = 1, 2, \dots, n$$

$$\text{or } \text{MVP}_j = \text{MFC}_j, \quad j = 1, 2, \dots, n$$

where MVP_j is the marginal value product of the j th input and MFC_j (marginal factor cost) is its factor price. Thus the firm will continue using each input until its marginal contribution to revenues just covers its costs.

3. Example

Consider the production function given by $y = 15x - .5x^2$. Now let the price of output be given by $p = 5$ and the price of the input be given by $w = 10$. The profit maximization problem can be written

$$\begin{aligned}
 p &= \max_x [5f(x) + 10x] \\
 &= \max_x [5(15x + 0.5x^2) + 10x] \\
 &= \max_x [65x + 2.5x^2]
 \end{aligned}$$

If we differentiate with respect to x we obtain

$$5(15 + x) + 10 = 0$$

$$\Upsilon \quad x = -13$$

C. Relationship between cost minimization and profit maximization

1. Duality of $V(y)$ and T

Since the cost function is defined by solving a minimization problem using the production technology, if we know the cost function for a firm, we can deduce the underlying technology. Thus we say the cost function and the technology are dual to each other, whether technology is represented by T or by $V(y)$.

2. Two-step optimization

Once the firm has determined the least costly way to produce each output level, it can maximize profits by choosing the level of output. Specifically, it has the following maximization problem

$$\max_y [py - C(y,w)]$$

If we differentiate this with respect to y we obtain

$$p - \frac{dC(y,w)}{dy} = 0$$

$$\Upsilon \quad p = MC$$

where MC is the cost of producing the last unit of output, or marginal cost.

Consider the following cost function for the firm

$$C(y) = 500 + 5y + 20y^2$$

Marginal cost is given by

$$C'(y) = 5 + 40y$$

If the price of output is 355, profit is given by

$$p = \max_y [355y - C(y)]$$

$$= \max_y [355y - (500 + 5y + 20y^2)]$$

If we differentiate this with respect to y we obtain

$$\begin{aligned} 355 & \frac{dC(y)}{dy} = 0 \\ Y & 355 + (5\% \cdot 40y) = 0 \\ Y & 355 = 5\% \cdot 40y \\ Y & 360 = 40y \\ Y & y = 9 \end{aligned}$$

III. Economies of size and scale

A. Economies of scale

1. Definitions

Consider the production function is given by

$$y = f(x_1, x_2, \dots, x_n) = f(x)$$

where y is output and x is the vector of inputs x_1, \dots, x_n . **The rate at which the amount of output, y , increases as all inputs are increased proportionately is called the degree of returns to scale for the production function $f(x)$.** The function f is said to exhibit **nonincreasing returns to scale** if for all $x \in R_+^n$, $\mu \geq 1$, and $\mu \neq 1$,

$$f(\mu x) \leq \mu f(x) \text{ and } \mu f(x) \neq f(\mu x)$$

Thus the function increases less than proportionately as all inputs x are increased in the same proportion, and it decreases less than proportionately as all x decrease in the same proportion. When inputs all increase by the same proportion we say that they increase along a ray.

In a similar fashion, we say that f exhibits **nondecreasing returns to scale** if for all $x \in R_+^n$, $\mu \geq 1$, and $0 < \mu \neq 1$

$$f(\mu x) \geq \mu f(x) \text{ and } \mu f(x) \neq f(\mu x)$$

The function f exhibits **constant returns to scale** if for all $x \in R_+^n$ and $\mu > 0$

$$f(\mu x) = \mu f(x)$$

This global definition of returns to scale is often supplemented by a local one that yields a specific numerical magnitude. This measure of returns to scale will be different depending on the levels of inputs and outputs at the point where it is measured. For one input, the **elasticity of scale** is

$$e = \frac{Mf(x)}{Mx} \frac{x}{f(x)} = \frac{My}{Mx} \frac{x}{y}$$

This is simply the elasticity of the marginal product of x . For multiple inputs, we have the more complicated formula:

$$e = \sum_{i=1}^n MP_i \frac{x_i}{y} = \sum_{i=1}^n \frac{Mf(x)}{Mx_i} \frac{x_i}{y}$$

Thus the returns to scale from increasing all of the inputs is the average marginal increase in output from all inputs, where each input is weighted by the relative size of that input compared to output. If e is less than one, then the technology is said to exhibit decreasing returns to scale; if it is equal to one, then the technology exhibits constant returns to scale; and if e is greater than one, the technology exhibits increasing returns to scale.

2. Implications

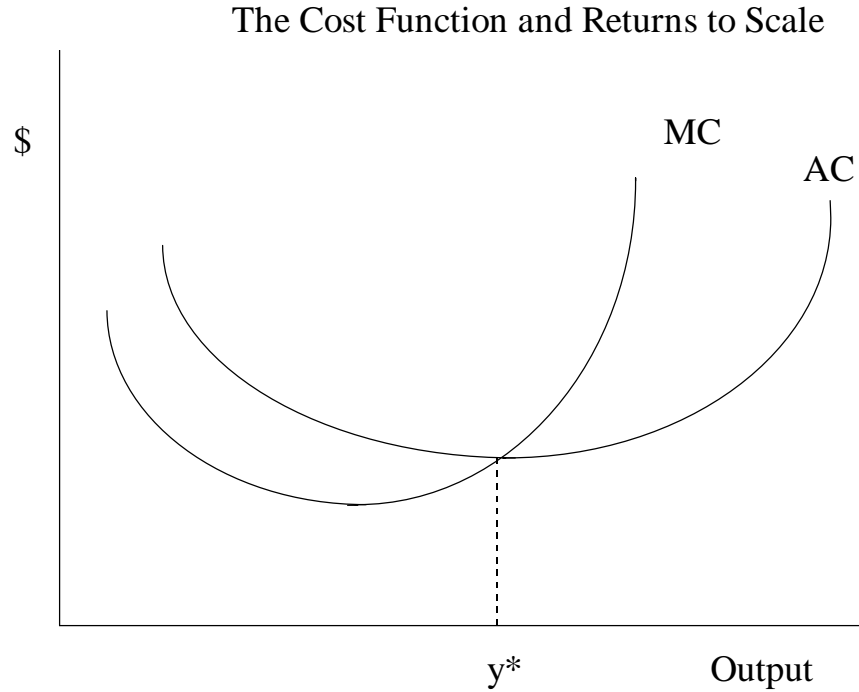
If a technology exhibits constant returns to scale then the firm can expand operations proportionately. If the firm can produce 5 units of output with a profit per unit of \$20, then by doubling the inputs and producing 10 units the firm will have a profit of \$40. Thus the firm can always make more profits by expanding. If the firm has increasing returns to scale, then by doubling inputs it will have more than double the output. Thus if it makes \$20 with 5 units it will make more than \$40 with 10 units etc. This assumes in all cases that the firm is increasing inputs in a proportional manner. If the firm can reduce the cost of an increased output by increasing inputs in a manner that is not proportional to the original inputs, then its increased economic returns may be larger than that implied by its scale coefficient. (See below.)

3. Returns to scale and the cost function

Given any output level and the associated cost-minimizing level of inputs we can consider measuring returns to scale at that point using the cost function since the elasticity of scale can be represented in terms of the parameters of the cost function. It is fairly easy to show using the conditions for cost minimization that

$$\begin{aligned}
 e &= \sum_{i=1}^n MP_i \frac{x_i}{y} \\
 &= \sum_{i=1}^n \frac{Mf(x)}{Mx_i} \frac{x_i}{y} \\
 &= \frac{\sum_{i=1}^n \frac{w_i}{Mc(w,y)} \frac{x_i}{y}}{My} \\
 &= \frac{\sum_{i=1}^n w_i x_i}{y \frac{Mc(w,y)}{My}} \\
 &= \frac{\sum_{i=1}^n w_i x_i}{y} \frac{1}{\frac{Mc(w,y)}{My}} \\
 &= \frac{AC}{MC}
 \end{aligned}$$

At cost-minimizing points, this cost ratio measure of elasticity of scale will be equivalent to the production function measure. The idea is that if AC is greater than MC then marginal cost is below average cost and so increasing output will lower average cost. Thus the firm can produce at lower average cost by expanding output.

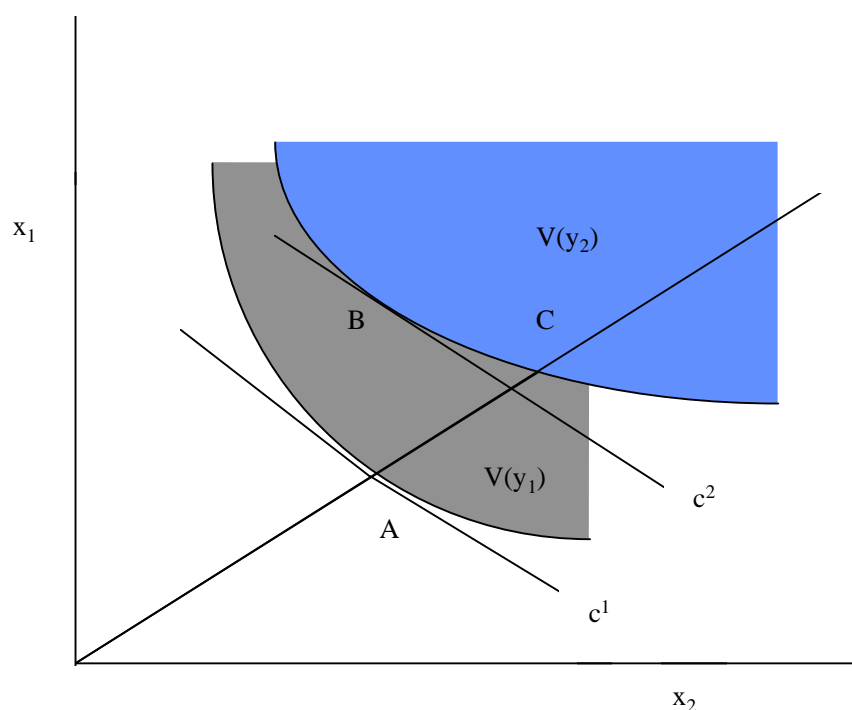


At output levels less than y^* , $AC > MC$ and $e > 1$ so there are increasing returns to scale. At output levels larger than y^* , $AC < MC$ and $e < 1$. At output y^* , $AC = MC$ and $e = 1$.

B. The expansion path of the firm and returns to scale

We discussed above the cost minimization problem for the firm. At a different level of output there will be a different cost-minimizing combination of inputs. Starting from one level of y , the set of all cost-minimizing input combinations as output expands is called the expansion path of the firm. This set of cost minimizing inputs may be proportional (contain points that are all proportional to each other), but it is likely that it is not. Graphically a proportional increase in inputs is represented by a straight line through the origin. Thus, starting from any cost minimizing input level, the firm can expand output by proportionately increasing inputs or by increasing them in a cost minimizing fashion. Returns to scale indicates the expansion in output from a proportional increase in inputs. Returns to size indicates the increase in output from a cost minimizing increase in inputs. For very small changes the two measures will be approximately equivalent.

Consider the following graphical representation of two output levels



Assume that for output level y_1 the cost-minimizing input combination is at point A. If the firm expands inputs proportionately enough to produce y_2 , it will be at point C. This point will have higher cost than at point B since it lies above the constant cost line c^2 .

C. Economies of size

1. Definition

The easiest way to find a construct a measure of economies of size is to use the **indirect production function** given by

$$IP(w, c^0) = \max_y [y \mid c(w, y) \leq c^0] = y(w, c^0)$$

This gives the maximum production for a given expenditure level denoted by c^0 . The returns to size considers how $IP(w, c^0)$ changes as the fixed expenditure level c^0 changes when input prices are held fixed. Thus, returns to size measures how maximal **output** changes as a fixed level of expenditure (**cost**) changes when input prices are held fixed. The idea is to see how much output will expand with a increase in expenditure. In a manner analogous to returns to scale, if maximal output increases proportionately more than cost (expenditure level), the technology exhibits increasing returns to size. If the maximal output increases proportionately less than cost, the technology exhibits decreasing returns to size. Finally, if they increase at proportionally the same rate, then the technology exhibits constant returns to size.

We can calculate returns to size as an elasticity using the indirect production function defined above:

$$\eta = \frac{\partial y(w, c^0)}{\partial c^0} \frac{c^0}{y}$$

Equivalently we can write it as the reciprocal of the elasticity of cost with respect to output,

$$\eta = \frac{1}{\left(\frac{\partial c(y, w)}{\partial y}\right) \left(\frac{y}{c}\right)} = \frac{c(y, w)}{\left(\frac{\partial c}{\partial y} y\right)} = \frac{AC}{MC}$$

Thus the elasticity of size is the ratio of average to marginal cost. If the elasticity of size (cost elasticity) is greater (less) than one, then the firm exhibits increasing returns to size and the average cost curve is downward sloping. If the average cost curve is upward sloping, then the size elasticity is less than one and the size elasticity is equal to one when average cost is at its minimum. Thus an examination of the shape of average costs curves provides an easy way to measure economies of size.

2. Examples

- a. Some example functional forms

$$\begin{aligned} C^1 &= 25 + 4y + .4y^2 + .03y^3 \\ C^2 &= 5 + .5y^{1.5} \\ C^3 &= .5y^{1.5} \end{aligned}$$

Notice that the third function is just the variable cost function for the second function.

- b. The marginal cost functions the above examples are, respectively,,

$$\begin{aligned} MC^1 &= 4 + .8y + .09y^2 \\ MC^2 &= .75y^{.5} \\ MC^3 &= .75y^{.5} \end{aligned}$$

The second and third functions have the same marginal cost.

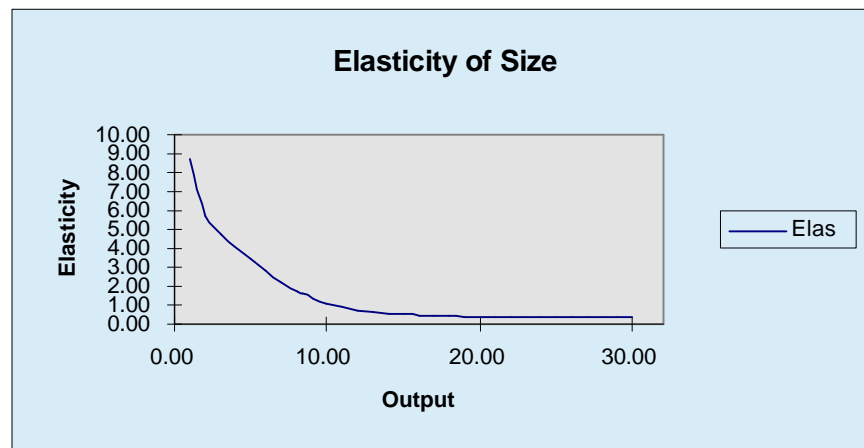
- c. Returns to size for each function

$$\begin{aligned} \eta^1 &= \frac{c(y,w)}{\left(\frac{Mc}{My}\right)y} = \frac{25 + 4y + .4y^2 + .03y^3}{(4 + .4y + .09y^2)y} \\ \eta^2 &= \frac{5 + .5y^{1.5}}{(.75y^{.5})y} = \frac{5 + .5y^{1.5}}{.75y^{1.5}} = \frac{5}{.75y^{1.5}} + \frac{.5y^{1.5}}{.75y^{1.5}} = 6.666y^{-1.5} + .666 \\ \eta^3 &= \frac{.5y^{1.5}}{(.75y^{.5})y} = \frac{.5y^{1.5}}{.75y^{1.5}} = \frac{.5}{.75} = .666 \end{aligned} \quad (56)$$

The first function has a variable elasticity of size. When y is equal to 5 the elasticity of size is equal to 3.44, which indicates increasing returns to size at this output level. At an output level of 10 the returns to size is 1.1 while at 15 it is .52. Thus this firm has first increasing and then decreasing returns to size. For the second function returns to size will fall as y increases while for the third one elasticity of size is constant.

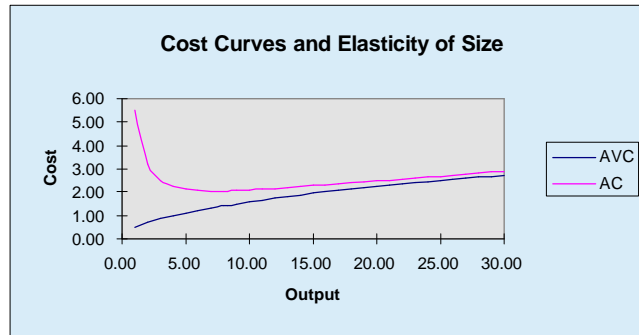
e. Graphical analysis

For the first cost function the elasticity of size looks as follows



Notice that the elasticity starts out greater than 1, is 1 between 10 and 11 units of output and then falls thereafter. This pattern is consistent with U-shaped average cost curves.

For the second and third functions the cost curves are as follows



Notice that average variable costs are always rising, so marginal cost is greater than average cost and the elasticity of size is always be less than 1. In the case of average total cost, for low output levels the curve is still falling indicating the elasticity is greater than 1.

D. Measurement of economies of size and scale

One can use a variety of empirical techniques to determine the extent of economies of scale for a given firm or group of firms within an industry. Some of these are positive in character, meaning they use data on firms in the industry; others are conditionally normative, meaning they involve the construction of representative firms or technologies using engineering, experimental or incomplete firm cost data. The most common positive technique is to estimate the cost functions using data on given firms or a cross section of firms within an industry. Common normative techniques include the construction of production budgets for specific products and the development of mathematical programming models that mimic the profit maximization or cost minimization problem for a synthetic or representative firm.

E. Empirical studies of size and scale

1. Empirical work on livestock production has generally found some economies of size in the production of feed cattle, pork and dairy. Factors contributing to economies of size include facilities such as fencing and flooring, equipment used in feed handling and mixing, and the use of specialized labor. The disposal of waste has differing effects. Large volumes of waste are sometimes difficult to get rid of, but there may be economies in the equipment or in other techniques (lagoons) used to handle them. These economies of size are consistent with the gradual concentration in the livestock industry.
2. Most studies of economies of size in crop production have found what is known as an L-shaped average cost curve. Economies of size occur as the firm expands from low output levels to more moderate ones but the costs become relatively flat beyond some level of output. Thus there are no great diseconomies of size for large firms but they have lower costs than the very small firms. Most studies of crop production have used normative methods, particularly linear programming. Most studies of cash grain production have found that costs per unit are falling for acreage below 500-750 acre, but do not drop much after that. The implication then is that small farms may have problems competing, but that medium size farms are competitive with much larger farms.

3. Studies of the processing industry indicate significant returns to size. Meat packing is a good example where larger firms have lower costs than smaller firms. The same holds in flour milling, soybean crushing or the production of corn sweetener. Thus, the number of firms in these industries is falling to a small and stable level.

IV. Multiproduct firms

A. Introduction

Most firms do not produce a single product, but rather, a number of related products. For example it is common for farms to produce two or more crops, such as corn and soybeans, barley and alfalfa hay, wheat and dry beans, etc. A flour miller may produce several types of flour and a food wholesaler carries a large number of products. A firm that produces several different products is called a multiproduct firm.

B. Multiproduct returns to scale

Consider the production possibility set of the multi-product firm

$$T = \{(x, y) \mid y \text{ can be produced by } x\}$$

where y and x are vectors of outputs and inputs, respectively.

We define the multiproduct elasticity of scale by

$$e_m = \sup\{r \mid \text{there exists a } d > 1 \text{ such that } (\lambda x, \lambda^r y) \in T \text{ for } 1 \leq \lambda \leq d\}$$

For our purposes we can regard the sup as a maximum. The constant of proportion λ is greater than or equal to 1. This gives the maximum proportional growth rate of outputs along a ray, as all inputs are expanded proportionally. (Baumol, Panzar, and Willig 1982). The idea is that we expand inputs by some proportion λ and see how much outputs can proportionately expand and still be in the production set. If $r = 1$, then we have constant returns to scale. If $r < 1$ then we have decreasing returns, etc.

C. Multiproduct returns to scale and the cost function

1. Definition

In the single product case, returns to scale is derived from the cost function as the ratio between cost and marginal cost times output; this reduces to average cost divided by marginal cost (AC/MC). In the multiproduct case, returns to scale is equal to the ratio of total cost to the sum of the marginal costs of each output multiplied by the output levels. Specifically returns to scale is given by

$$e_m = \frac{c(w, y)}{\sum_{j=1}^m \frac{Mc(w, y)}{y_j} y_j}$$

The marginal cost associated with each output, holding the others fixed, is weighted by the associated output level. This cost function description of returns to scale is the same as the one computed directly using the technology set T as long as we consider only cost-minimizing points on the expansion path.

2. Ray average costs

There is no meaningful way to define average costs for a multiproduct firm. It would make no sense to divide total costs by the level of any one output. The point is that there are several outputs and no natural divisor exists in order to compute average cost. Instead, we use a composite good made up of all outputs y_j to measure average cost. We consider average cost for different levels of this composite good, along a ray in output space, since proportionate increases in $y_1 \dots y_n$ lead to similar increases in the composite good. The idea is to measure costs as outputs expand in a constant proportion. Consider then a particular output level y^0 which is considered as base combination of outputs. Average cost at the point y along a ray through y^0 is called ray average costs and is defined by

$$RAC = \frac{c(\lambda y^0, w)}{\lambda}$$

The factor λ is the units of the base bundle of outputs y^0 in the bundle $y = \lambda y^0$. For example, if the y we are considering is $(2,3)$, then other bundles on the same ray are $(4,6)$, $(8,12)$ $(2/3, 1)$ and $(2/5, 3/5)$.

The question then is which bundle we choose as the one which is assigned the value $\lambda=1$ along the ray. The convention (Baumol, Panzar and Willig 1982) is to use the bundle whose components sum to one as the base. For the example above, the base vector is $y^0 = (2/5, 3/5)$. All other outputs on the ray are scalar multiples of $(2/5, 3/5)$. For example the output combination $(2,3) = 5(2/5, 3/5)$.

So we can express every y along the ray through y^0 as $y = \lambda y^0$ for some λ , different for each y . For the combination $(2,3)$, $\lambda = 5$, for the combination $(8,12)$, $\lambda = 20$ and for $(2/5, 3/5)$, $\lambda = 1$. In general $\lambda = \sum y_j$. Then ray average cost at the point y (written as λy^0) is given by

$$RAC = \frac{c(\lambda y^0, w)}{\lambda} = \frac{c\left(\sum_{j=1}^m y_j y^0, w\right)}{\sum_{j=1}^m y_j} = \frac{c(y, w)}{\sum_{j=1}^m y_j}$$

The use of the sum of the outputs as a divisor may smack of adding apples and oranges, but the weighting is truly arbitrary since all changes are measured along a ray.

The elasticity of ray average cost is given by

$$e = \frac{\% \Delta RAC(y)}{\% \Delta y} = \frac{\% \Delta RAC(y)}{\% \Delta y}$$

The elasticity of ray average cost is related to scale economies and it can be shown that (Baumol, Panzar, and Willig 1982, p. 51)

$$\% \Delta RAC(y) = \frac{1}{\% e}$$

Thus returns to scale is greater (less) than one whenever the elasticity of ray average cost is less (greater) than zero, or equivalently, whenever average cost is decreasing (increasing). As we might expect, constant returns to scale is equivalent to a constant ray average cost.

3. Multiproduct returns to size

In the single product case, returns to size, defined as the reciprocal of the cost function elasticity, measures changes in costs as inputs expand in a cost minimizing non-proportional manner with increases in output. No similar measure exists in the case of the multiproduct firm since there is not a natural way to expand multiple outputs other than proportionately, and this gives multiproduct returns to scale as discussed above.

4. Product-specific returns to scale and incremental costs

While the overall returns to scale provides information on economies as the firm expands, it is also useful to study changes in cost as individual products expand. To analyze such changes, the concept of product-specific returns to scale is used. This is related to the incremental cost of producing a particular good. Consider a firm producing m outputs and denote this set of outputs by M . Let all outputs except the i th be denoted by $M-i$. The notation y_{M-i} denotes the vector $(y_1, y_2, \dots, y_{i-1}, 0, \dots, y_m)$, where we set the i th element to 0. Thus $c(y_{M-i}) = c(y_1, y_2, \dots, y_{i-1}, 0, \dots, y_m)$.

The incremental cost of producing the i th good in quantity y_i is given by

$$IC_i(y) = c(y) - c(y_{M-i})$$

and the average incremental cost is given by:

$$AIC_i = \frac{IC_i(y)}{y_i}$$

For the two output case we obtain

$$IC_1(y_1, y_2) = c(y_1, y_2) - c(0, y_2)$$

$$AIC_1 = \frac{IC_1(y_1, y_2)}{y_1}$$

Product specific returns to scale is then defined to be the ratio between incremental cost and the

product of marginal cost and output y_i . I.e.,

$$PSE_i = \rho_i = \frac{IC_i}{y_i} \cdot \frac{AIC_i}{MC_i}$$

much the same as returns to scale is the ratio between cost and marginal cost times output. When ρ_i is greater than one, we have increasing returns to scale with respect to y_i . In a similar fashion, returns to scale for a group of products can be defined by evaluating the incremental cost of increasing these products from zero to some level holding all other products constant, i.e., for a set L which is a subset of M

$$PSE_L = \rho_L = \frac{IC_L(y)}{\sum_{j \in L} \frac{MC}{y_j} y_j}$$

For example L might be crops and M-L might be livestock. Thus ρ would be the returns to scale of crops, as a part of the whole enterprise. If we divide the different outputs into groups in this way, we can find the overall measure of returns to scale as a weighted average of the group specific ones. Specifically if the subscript L denotes a subset of the outputs and M denotes all outputs, then

$$\rho_m = \frac{a_L \rho_L + (1-a_L) \rho_{M\&L}}{\frac{IC_L + IC_{M\&L}}{c}} \quad \text{where}$$

$$a_L = \frac{\sum_{j \in L} y_j \frac{MC}{y_j}}{\sum_{j \in M} y_j \frac{MC}{y_j}}$$

Note that if the denominator in the last expression were unity, then overall returns to scale would be the weighted average of those of the groups. If the production process is nonjoint so that the outputs do not affect each other, the denominator will be one; otherwise, overall returns to scale depends on interactions between the outputs, as would be the case with corn and soybeans, alfalfa and barley, or corn sweetener and corn gluten.

D. Economies of scope

1. Idea

While economies of scale or size explain cost changes that occur as output expands, costs may also change due to changes in the product mix. If there are cost advantages of producing several products simultaneously, rather than separately, then economies of scope are said to exist.

2. Definition

A formal definition requires the division of the outputs into a number of non-overlapping groups. We will denote the set of all products as M . For the one Iowa farmer M might be as follows: $M = (\text{corn, soybeans, feeder pigs, market hogs})$. Let S be a subset of M such as corn and soybeans. Then let P be a non-overlapping partition of S , with each element denoted L . The union of the LOP is equal to the set S and the sets L and L_N are nonintersecting for $L \dots L_N$. Then economies of scope are said to exist for partition P if

$$c(y_S) < \sum_{LOP} c(y_L)$$

where for each set L , $c(y_L)$ is the cost of producing only the outputs in L , with all other outputs set to zero.

3. Example with two outputs

If there are only two outputs then $M = \{1, 2\}$. Consider $S = \{1, 2\}$. The only reasonable division of S is $(\{1\}, \{2\})$ where L_1 is the first output and L_2 the second output. In this case economies of scope exist if

$$C(y_1, y_2) < C(y_1, 0) + C(0, y_2)$$

The sum of the costs of producing the goods separately is higher than the cost of producing them together.

4. Example with more than two outputs

Now consider an case where there are more than 2 outputs. For example, if $M = \{\text{corn, soybean, feeder pigs, market hogs}\}$, we might consider whether there are economies of scope in growing the corn and soybean. In this case we define $S = \{\text{corn, soybean}\}$ and $P = \{\{\text{corn}\}, \{\text{soybean}\}\}$. Economies of scope exist when

$$c(\text{corn, soybean}) < c(\text{corn}, 0) + c(0, \text{soybean})$$

This is probably the case since it is less expensive to grow the same amounts of corn and soybean together than to grow these amounts separately. This occurs because of the nitrogen fixation properties of soybean.

Now let $S = M = (\text{corn, soybeans, feeder pigs, market hogs})$ and $P = [\{\text{corn, soybeans}\} \{\text{feeder pigs}\} \{\text{market hogs}\}]$. Economies of scope exist with respect to this partition P if

$$c(\text{corn, soybeans, feeder pigs, market hogs}) < c(\text{corn, soybeans, 0, 0}) \\ + c(0, 0, \text{feeder pigs}, 0) \\ + c(0, 0, 0, \text{market hogs})$$

If this is the case, then it is less expensive to produce feeder pigs and market hogs along with corn and soybeans than to produce all of these groups of products separately.

5. Measurement

It is possible to define the returns to scope for any partition P , by comparing the costs of producing product groups together and separately. Consider first the special, simple case, in which $S=M$ and P takes the form $P = \{L, M-L\}$. Thus, we divide M into two parts and consider the cost savings from producing these two parts together. The **returns to scope** is given by

$$SC_L(y) = \frac{[c(y_L) + c(y_{M\&L}) - c(y)]}{c(y)}$$

This is the difference between the cost of producing groups L and $M-L$ separately, and producing them together, divided by $c(y)$ as a normalization. The higher the difference in costs, the greater the savings from joint production.

When there are only two outputs, the returns to scope is simply:

$$SC_L(y_1, y_2) = \frac{C(y_1, 0) + C(0, y_2) - C(y_1, y_2)}{C(y_1, y_2)}$$

More generally, for a partition P , the returns to scope is given by

$$SC_P(y) = \frac{c(y_L) + c(y_S) - c(y)}{c(y)}$$

6. Relationship between economies of scope and multi-product returns to scale

Consider the simple case in which $S = M$ and $P = \{L, M-L\}$. Overall returns to scale is related to returns to scope and product specific returns to scale through the identity:

$$r_m = \frac{a_L r_L + (1 - a_L) r_{M\&L}}{1 + SC_L}$$

where ρ_L is product-specific returns to scale for group L, ρ_{M-L} is product-specific returns to scale for the rest of the products (M-L) and α is a weighting factor. If the returns to scope is zero, then overall returns to scale is just the weighted average of the group returns. If the returns to scope is non-zero then an adjustment must be made. For example, returns to scale might be higher when a producer increases corn and soybean acreage simultaneously rather than just one or the other alone.

7. Public inputs and economies of scope

One important source of economies of scope is the presence of public inputs. **These are inputs which can be used for one production process without reducing the amount available for other processes.** An example might be the use of ponds both for fish culture and to water for grazing animals. While perfect public inputs are rare, quasi-public inputs are very common. **Quasi-public inputs are those which can be shared by two production processes without complete congestion in use.** The use of the same planting equipment for corn and soybeans is a typical example since the demands of the two crops do not typically occur simultaneously. Many allocated fixed inputs in agriculture may lead to economies of scope. In agribusiness firms, economies of scope may exist in sales and advertising, since sales efforts can promote more than one product or line.

E. Examples

1. Functional form and example points

Let the multiproduct cost function be given by

$$cost = .25y_1^2 + y_2 + y_1^2 + 2y_1y_2 + .5y_2^2 + .1y_1^3$$

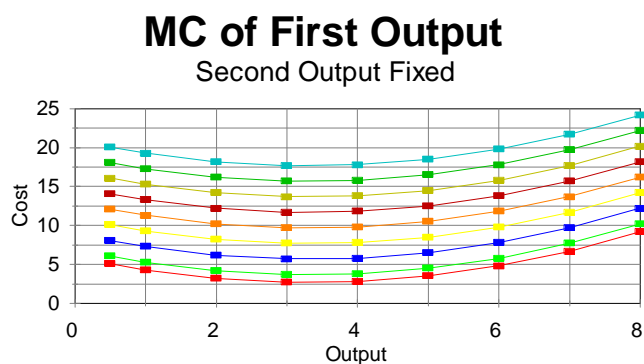
2. Marginal cost

Marginal cost is given by the derivative of the cost function with respect to the respective outputs.

$$MC_1 = .5 + 2y_1 + 2y_2 + .3y_1^2$$

$$MC_2 = .1 + 2y_1 + y_2$$

Data on the marginal cost of y_1 and y_2 for a variety of output combinations are contained in Table 1. If we hold y_2 fixed at various levels we get a family of marginal cost curves for y_1 .



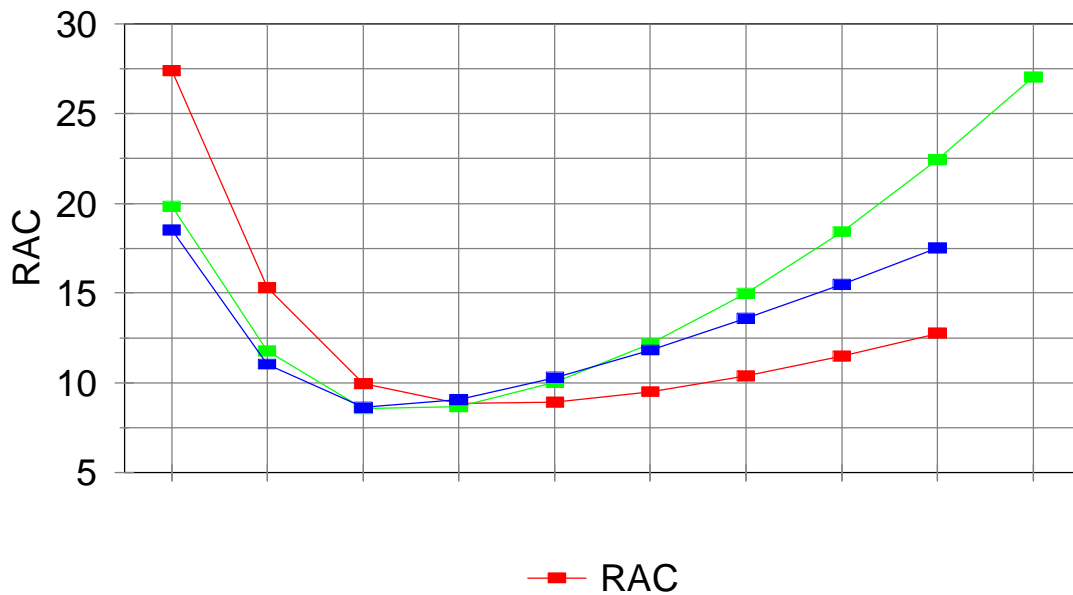
3. Ray average cost

$$RAC(y_1, y_2) = \frac{0.25y_1 + y_2 + y_1^2 + 2y_1y_2 + 0.5y_2^2 + 0.1y_1^3}{y_1 + y_2}$$

Ray average cost is given by $C(y, w)/S y_j$. For this case we obtain

Data on ray average cost are contained in Table 1. Below are ray average cost curves for two rays. Notice that both curves are U-shaped. RAC falls at first, and then eventually rises, consistent with initial increasing returns to scale.

Ray Average Costs



4. Economies of scale from the cost function

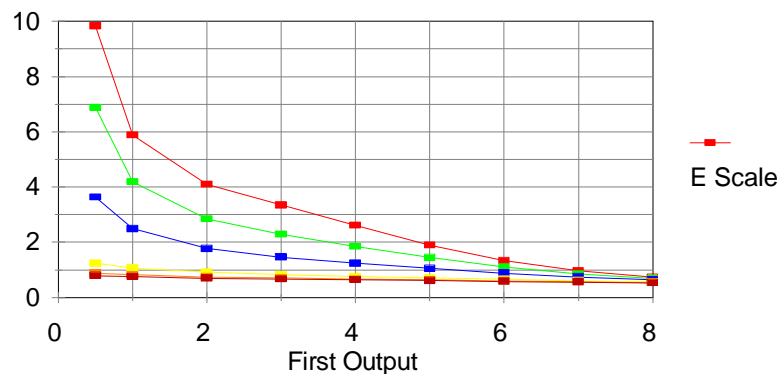
Multiproduct returns to scale is given by

$$\rho_m(y_1, y_2) = \frac{c(w, y)}{\sum_{j=1}^m \frac{\partial c(w, y)}{\partial y_j} y_j}$$

$$= \frac{25\%5y_1 + y_2 + y_1^2\%2y_1y_2 + .5y_2^2 + .1y_1^3}{(5 + 2y_1 + 2y_2 + .3y_1^2)y_1 + (1 + 2y_1 + y_2)y_2}$$

We can examine this for various levels of y_1 holding y_2 constant as we trace out the curve. The top curve holds y_2 constant at $y_2 = 5$. The bottom curve holds $y_2 = 8$.

Economies of Scale



Notice that ρ_m is greater than one when both output levels are low, but falls to less than one as y_1 increases.

5. Average incremental cost for output y_i

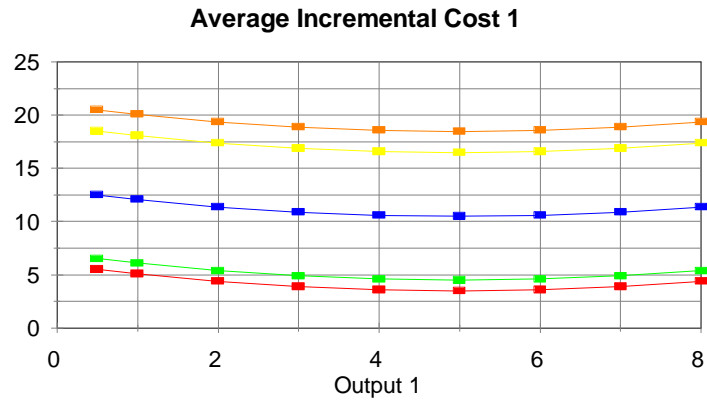
$$AIC_i(y) = \frac{c(y) - c(y_{-i})}{y_i}$$

For the two outputs y_1 and y_2 we obtain

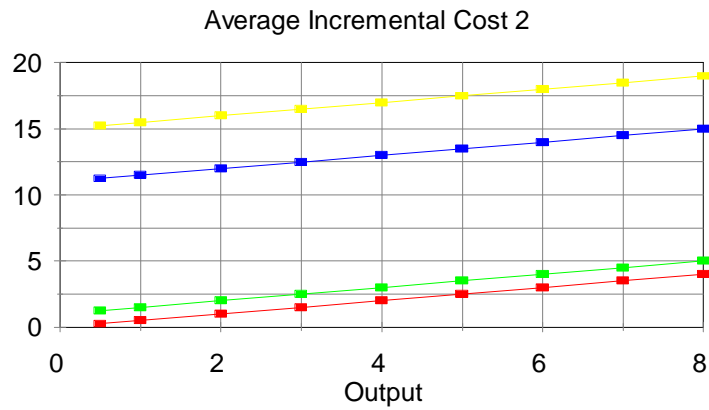
$$AIC(y_1) = \frac{(25\%5y_1 + y_2 + y_1^2\%2y_1y_2 + .5y_2^2 + .1y_1^3) - (25\%5y_2 + .5y_2^2)}{y_1}$$

$$AIC(y_2) = \frac{(25\%5y_1 + y_2 + y_1^2\%2y_1y_2 + .5y_2^2 + .1y_1^3) - (25\%5y_1 + y_1^2\%2y_1y_2 + .1y_1^3)}{y_2}$$

The average incremental costs of y_1 are as follows (for different fixed levels of y_2)



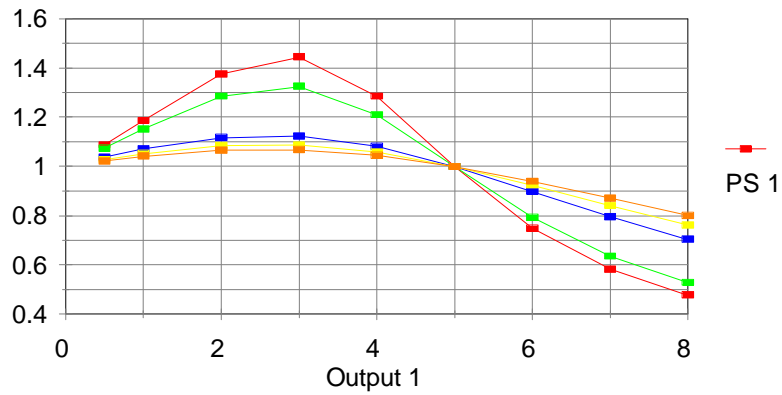
while for y_2 they are always upward sloping as follows (for different fixed levels of y_1)



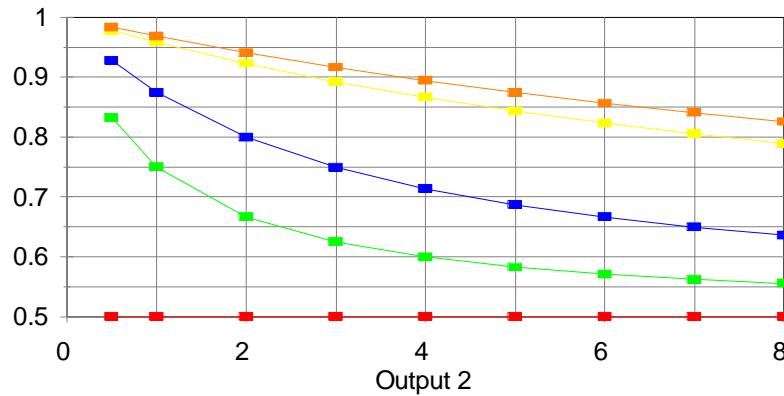
6. Product-specific economies of scale

Product-specific returns to scale is given by dividing AIC by MC. (see section IV.C.4) For this example, for y_1 , this is greater than one at low output levels and then falls as y_1 increases (for all levels of y_2). For y_2 , product-specific returns to scale is always less than one.

Product Specific Economies of Scale 1



Product Specific Economies of Scale 2



7. Economies of scope

Returns to scope for the partition $\{L, M-L\}$ is given by

$$SC_L(y) = \frac{c(y_L) + c(y_{M-L}) - c(y)}{c(y)}$$

For this example, this is given by

$$SC(y_1, y_2) = \frac{(25y_1 + y_1^2 + 1y_1^3) + (25y_2 + 5y_2^2) - (25y_1 + y_2 + y_1^2 + 2y_1y_2 + 5y_2^2 + 1y_1^3)}{(25y_1 + y_2 + y_1^2 + 2y_1y_2 + 5y_2^2 + 1y_1^3)}$$

The following graph plots returns to scope as y_1 varies, holding y_2 constant at various levels. Notice that it falls below zero for higher output levels.

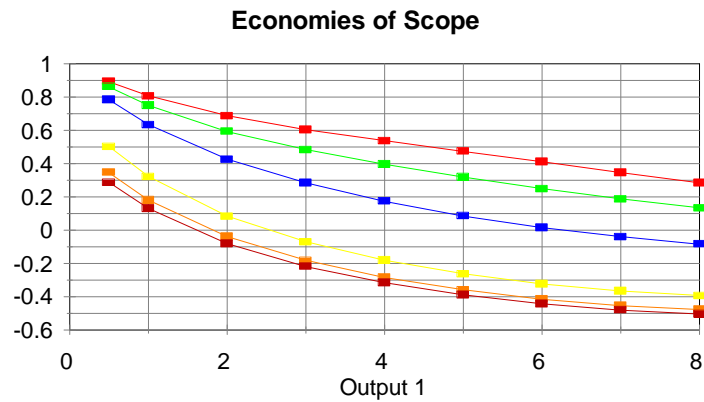


Table 1 Multiproduct Cost Data

y1	y2	Cost	MC1	MC2	RAC	E Scale	C - 1	AIC 1	C - 2	AIC 2	PS 1	PS 2	Scope
0.5	0.5	27.388	5.075	0.500	27.388	9.825	24.63	5.525	27.263	0.250	1.089	0.500	0.895
1.0	0.5	29.725	4.300	1.500	19.817	5.886	24.63	5.100	29.100	1.250	1.186	0.833	0.807
2.0	0.5	33.425	3.200	3.500	13.370	4.101	24.63	4.400	31.800	3.250	1.375	0.929	0.688
3.0	0.5	36.325	2.700	5.500	10.379	3.348	24.63	3.900	33.700	5.250	1.444	0.955	0.606
4.0	0.5	39.025	2.800	7.500	8.672	2.610	24.63	3.600	35.400	7.250	1.286	0.967	0.538
5.0	0.5	42.125	3.500	9.500	7.659	1.893	24.63	3.500	37.500	9.250	1.000	0.974	0.475
6.0	0.5	46.225	4.800	11.500	7.112	1.338	24.63	3.600	40.600	11.250	0.750	0.978	0.411
7.0	0.5	51.925	6.700	13.500	6.923	0.968	24.63	3.900	45.300	13.250	0.582	0.981	0.347
8.0	0.5	59.825	9.200	15.500	7.038	0.735	24.63	4.400	52.200	15.250	0.478	0.984	0.284
0.5	1.0	27.763	6.075	1.000	18.508	6.876	24.5	6.525	27.263	0.500	1.074	0.500	0.864
1.0	1.0	30.600	5.300	2.000	15.300	4.192	24.5	6.100	29.100	1.500	1.151	0.750	0.752
2.0	1.0	35.300	4.200	4.000	11.767	2.847	24.5	5.400	31.800	3.500	1.286	0.875	0.595
3.0	1.0	39.200	3.700	6.000	9.800	2.292	24.5	4.900	33.700	5.500	1.324	0.917	0.485
4.0	1.0	42.900	3.800	8.000	8.580	1.849	24.5	4.600	35.400	7.500	1.211	0.938	0.396
5.0	1.0	47.000	4.500	10.000	7.833	1.446	24.5	4.500	37.500	9.500	1.000	0.950	0.319
6.0	1.0	52.100	5.800	12.000	7.443	1.113	24.5	4.600	40.600	11.500	0.793	0.958	0.250
7.0	1.0	58.800	7.700	14.000	7.350	0.866	24.5	4.900	45.300	13.500	0.636	0.964	0.187
8.0	1.0	67.700	10.200	16.000	7.522	0.694	24.5	5.400	52.200	15.500	0.529	0.969	0.133
0.5	2.0	29.263	8.075	2.000	11.705	3.641	25	8.525	27.263	1.000	1.056	0.500	0.786
1.0	2.0	33.100	7.300	3.000	11.033	2.489	25	8.100	29.100	2.000	1.110	0.667	0.634
2.0	2.0	39.800	6.200	5.000	9.950	1.777	25	7.400	31.800	4.000	1.194	0.800	0.427
3.0	2.0	45.700	5.700	7.000	9.140	1.469	25	6.900	33.700	6.000	1.211	0.857	0.284
4.0	2.0	51.400	5.800	9.000	8.567	1.248	25	6.600	35.400	8.000	1.138	0.889	0.175
5.0	2.0	57.500	6.500	11.000	8.214	1.055	25	6.500	37.500	10.000	1.000	0.909	0.087
6.0	2.0	64.600	7.800	13.000	8.075	0.887	25	6.600	40.600	12.000	0.846	0.923	0.015
7.0	2.0	73.300	9.700	15.000	8.144	0.749	25	6.900	45.300	14.000	0.711	0.933	-0.041
8.0	2.0	84.200	12.200	17.000	8.420	0.640	25	7.400	52.200	16.000	0.607	0.941	-0.083
0.5	3.0	31.763	10.075	3.000	9.075	2.263	26.5	10.525	27.263	1.500	1.045	0.500	0.693
1.0	3.0	36.600	9.300	4.000	9.150	1.718	26.5	10.100	29.100	2.500	1.086	0.625	0.519
2.0	3.0	45.300	8.200	6.000	9.060	1.317	26.5	9.400	31.800	4.500	1.146	0.750	0.287
3.0	3.0	53.200	7.700	8.000	8.867	1.130	26.5	8.900	33.700	6.500	1.156	0.813	0.132
4.0	3.0	60.900	7.800	10.000	8.700	0.995	26.5	8.600	35.400	8.500	1.103	0.850	0.016
5.0	3.0	69.000	8.500	12.000	8.625	0.879	26.5	8.500	37.500	10.500	1.000	0.875	-0.072
6.0	3.0	78.100	9.800	14.000	8.678	0.775	26.5	8.600	40.600	12.500	0.878	0.893	-0.141
7.0	3.0	88.800	11.700	16.000	8.880	0.684	26.5	8.900	45.300	14.500	0.761	0.906	-0.191
8.0	3.0	101.70	14.200	18.000	9.245	0.607	26.5	9.400	52.200	16.500	0.662	0.917	-0.226
0.5	4.0	35.263	12.075	4.000	7.836	1.600	29	12.525	27.263	2.000	1.037	0.500	0.596
1.0	4.0	41.100	11.300	5.000	8.220	1.313	29	12.100	29.100	3.000	1.071	0.600	0.414
2.0	4.0	51.800	10.200	7.000	8.633	1.070	29	11.400	31.800	5.000	1.118	0.714	0.174
3.0	4.0	61.700	9.700	9.000	8.814	0.948	29	10.900	33.700	7.000	1.124	0.778	0.016
4.0	4.0	71.400	9.800	11.000	8.925	0.858	29	10.600	35.400	9.000	1.082	0.818	-0.098
5.0	4.0	81.500	10.500	13.000	9.056	0.780	29	10.500	37.500	11.000	1.000	0.846	-0.184
6.0	4.0	92.600	11.800	15.000	9.260	0.708	29	10.600	40.600	13.000	0.898	0.867	-0.248
7.0	4.0	105.30	13.700	17.000	9.573	0.642	29	10.900	45.300	15.000	0.796	0.882	-0.294
8.0	4.0	120.20	16.200	19.000	10.017	0.585	29	11.400	52.200	17.000	0.704	0.895	-0.324

V. Beyond the Neoclassical Firm

In the discussion above, a firm is simply a production function. The objective of the firm is to maximize profits. In reality, firms must deal with many complex human challenges, such as creating incentives, and coping with incomplete information.

A. Agency Issues: Labor is not a simple factor of production

The production function model of a firm views labor as an input like any other. However, labor is different, since the workers must be motivated to work effectively (supply the input purchased). Supplying effective incentives may be difficult, because the employer cannot have complete information about the effort a worker is exerting. **Agency costs** refer to the costs an employer incurs in motivating workers, or the loss in efficiency which arises because monitoring is imperfect.

1. Monitoring Costs

One aspect of agency costs is **monitoring costs**. When a worker's productivity is difficult to observe, a firm must invest resources in making productivity easier to measure.

For example the individual output of a worker on an assembly line may be difficult to observe. Thus, in the absence of monitoring, a firm will find no alternative to paying workers a fixed wage per hour, which provides little incentive to workers to exert high effort. To alleviate this problem, and to allow for stronger incentives, the firm may hire a monitor, such as a foreman. His job, among other things, is to make sure that the workers do their jobs correctly and do not shirk on the job or do their tasks badly. He may do so either by providing positive encouragement to workers who do well (the carrot approach), or by penalizing workers who are not working properly (the stick).

The foreman himself must also be provided with incentives. Thus, organizations contain hierarchies. Each level monitors the next lowest level. The larger the firm, the more levels will be necessary, and the costs may be substantial. The higher monitoring costs are, the less the firm will engage in monitoring, and the lower incentives will be. A division within a large, vertically integrated firm is less likely to innovate than a small entrepreneurial firm, because few of the gains from innovation are likely to accrue to that division or its workers (because it is difficult to measure who is responsible for the gains). Only if the vertically integrated firm provides appropriate incentives will much innovation occur, and providing such incentives may be extremely costly, or even impossible.

Rather than investing in monitoring technologies, a firm may modify the production process to enhance the observability of worker effort and make monitoring more effective. This change in technology may actually be technologically inefficient, but it may still be economically efficient, since the cost of monitoring must be taken into account as well as the cost of the labor being monitored. For example, during the slave period, plantation owners issued short-handled hoes, not because they are more efficient, but because it was easier to monitor workers who are bent-over hoeing as opposed to standing.

2. Motivating workers

Because monitoring is costly, firms will monitor imperfectly, and they will look for incentive schemes which base compensation on numbers which are easy (inexpensive) to measure. Each strategy has its pitfalls, but may be optimal under many circumstances.

- a. A car manufacturer who owns dealerships may find it difficult to ensure that the dealers are working hard to sell a large number of cars at good prices since the manufacturer has little

information about local market conditions. In this case, effective monitoring may be prohibitively costly. One solution is to create franchised dealers. A **franchise** is an independent business, authorized by the manufacturer to sell its goods. The owner receives the profits, less a royalty on sales. Thus, incentives are high. However, franchising is not free of problems: A franchisee faces a great deal of risk, and must be compensated for this risk in the form of lower franchise fees or royalties. Also, franchisees may emphasize profits at the expense of the brand's reputation. When these problems are particularly severe, company stores may be superior, even if incentives are weaker.

- b. Rather than paying workers a fixed wage a firm may choose to use **piece rates**. Examples include paying farm workers per bushel picked or row weeded. In a sewing plant, the pay might be per item completed. However, a worker who receives a piece rate has no incentive to keep quality high. For example, a worker may pick many cherries per hour, but he may smash half of them. Thus monitoring of some kind will be necessary, and the higher the emphasis on piece rates, the more monitoring will be needed. The higher the incentive for quantity, the greater the need for monitoring quality. Piece rates will be chosen when quality is relatively unimportant or easy to monitor. Otherwise, a simple hourly wage may be superior.
- c. A firm may pay wages higher than the market (or at least higher than necessary to retain the workers during a recession) in order to improve morale and enhance productivity. Such a wage is called an **efficiency wage**. If being fired is very costly to worker, she will be particularly careful to do a good job, even if the chance of being caught shirking on the job is small. The cost of being fired is higher when one's wage is higher. Therefore, the higher the wage, the less monitoring is necessary to obtain the same effort level from workers. The firm may actually save money by paying higher wages, since the savings on monitoring costs may be substantial. Some macroeconomists argue that efficiency wages are responsible for unemployment, since they keep wages above the level at which all workers can find jobs (Akerlof & Yellen 1985; Weiss, 1990). The higher wages are, the fewer workers employers are willing to hire.
- d. Firms may pay **bonuses based on group performance** if individual output is hard to observe. One common contest is the **contest** to reach top management levels with their disproportionately higher salaries. Even if the output of these workers does not match this salary level the motivational effect may pay off. The effectiveness of such bonuses is limited, since one cannot compensate every employee for the full increase in performance of the group. (There isn't enough money to go round.) A similar strategy is employee **stock ownership** or **profit-sharing**. Again, its effectiveness is limited, but it may improve morale by making workers see their interest as being more aligned with their employer's. Another example is the case of broiler tournaments where individuals contracting to produce broilers are paid a premium based on their performance relative to others who have signed contracts.

B. Defining the vertical boundaries of a firm

The vertical chain involves a large variety of specialized activities, ranging from product design and materials procurement to distribution and sales. A firm must decide which activities will take place within its boundaries, and which will take place outside, through relationships with other firms.

1. The make or buy decision (up-stream vertical integration)

A firm must decide how it will obtain the inputs (goods and services) that are needed in production. If the firm purchases inputs from another firm, then we say the firm is **using the market**. If the firm

chooses to provide a good or service for itself, then we say that it is **vertically integrated** with respect to that good or service. Thus the firm can choose whether to **make or buy** the inputs it needs in production. For example, a milk products company may decide to run its own dairies, or it may contract with existing dairies to obtain the milk necessary for producing such products as butter and yogurt. On the services side, a firm may have accountants on its permanent staff; but commonly, the firm will instead contract with an independent accounting firm for such services as tax form preparation and internal auditing. *What determines the choice of such a firm?*

2. Down-stream integration

Similarly, on the down-stream side, a firm may market products itself (by setting up sales locations and hiring sales personnel), or it may choose to use **independent contractors** or retailers to do this marketing. Automobile companies rely primarily on independent dealers (franchisees) to market their products. A food distributor rarely owns all of the supermarkets which carry its products. However, often large food products companies will operate their own distribution systems. So this industry is partially down-ward integrated, but not entirely so. *What determines the degree of down-stream integration?*

C. Neoclassical approaches to vertical integration

The basic neoclassical answer to the vertical integration question is: A firm should buy an input externally if the costs (and price offered) of the external producer are lower than the cost of producing the product internally. A firm should market its goods through an external distributor if the cost of marketing a product itself exceed the costs of contracting with such a distributor.

1. Misleading answers

In making such a determination, one must be careful to consider the *opportunity costs* associated with each choice. The following are misleading reasons often advanced for performing an activity internally, rather than externally:

- a. "Firms should make rather than buy in order to avoid paying a profit margin to independent firms." or "Firms should market products themselves in order that they can earn the entire profit earned on sales."

This is misleading since the firm will not find it worthwhile to produce or market a product itself unless some profit (or extra profit) can be earned. If markets are competitive, the profit earned by an independent firm is simply sufficient to make its activities worthwhile to the owner. For example, although a food distributor may earn a mark-up on the products it markets, it also must incur inventory costs. Unless the product manufacturer can distribute more efficiently than an external firm (i.e. with lower inventory costs), there is no reason to engage in distribution internally.

- b. "A firm should make rather than buy in order to avoid high market prices during periods of peak demand or scarce supply."

This makes no sense if we consider opportunity costs, since the firm must forego the high market price available in external sales if it uses an input internally rather than selling it on the market.

2. A legitimate reason to use the market: economies of size or scale

If economies of scale exist in the production of a product, then an independent firm, who can service

a large number of customers, may be better able to realize these economies than an individual firm producing only for its own use. As an example, consider trucking of livestock to a market 90 miles from the place of production. There may be economies of scale in using a large tractor-trailer to haul the livestock. However, if the producer ships only one truck load every two months, it may not be efficient for the firm to purchase a truck. In such a case, the use of custom trucking firm may be a reasonable use of the market. On the other hand, a large firm shipping a 2 loads of livestock 4 days a week may prefer to own its own truck.

Of course, the independent firm must pass the savings due to economies of size on to the buyer, or else there will be no incentive to use the market. As long as the market for this input is competitive, the independent firm will have every incentive to pass on the savings in order to gain business and make a profit.

Economies of scale may exist not only within the technology itself but also due to the organization of production. A firm that specializes in a particular product and services a large customer base may develop efficient ways to provide the product that would not be available to a firm making only a small quantity of the item.

D. Why is it ever better not to use the market? Beyond neoclassical answers.

Above, we found some reasons why a firm is better off using the market, and also some reasons why the firm *might* be able to do *as well* producing inputs or market outputs internally (being vertically integrated). But why is it that, in many circumstances, a firm actually *needs to be vertically integrated*. What is the problem with using the market?

To answer this question, we must go beyond the neoclassical model of production (which simply compares the costs of producing/distributing internally or externally). In order to explain why markets fail, we need to consider more complicated factors, such as transaction costs, incomplete contracts, and asymmetric information between market players.

1. The importance of coordination

Proper coordination of the various steps in production is essential to minimizing the costs of moving a product through the vertical chain. Suppliers must produce adequate amounts of the right quality and design. Distributors must be able to transport and warehouse the goods. Retailers must have appropriate space and the effective marketing concepts. Lack of coordination can effectively shut down or seriously reduce the effectiveness of a given segment of the chain. Consider the problem of the meat packer who is unable to obtain the right number and type of cattle for a given days kill.

Coordination is especially important when a product has specific design attributes that must fit together. A combine manufacturer may require a specific type of hydro-transmission for its new harvester line. This transmission must fit in the appropriate place, have the required number of hose fittings, be able to move the combine at specific velocities with certain torque, arrive in sufficient quantity at the right time, etc.

If the inputs for production are produced in-house, proper coordination is easy. Headquarters simply tells the different divisions what inputs are needed when, and in what quantities, and the end product is produced from these inputs.

2. Incomplete contracts: a result of bounded rationality and uncertainty

When using the market, a contract is often required to guarantee proper coordination, so that, for

example, inputs fit production needs. Unfortunately, it is difficult to specify, in a contract, precisely what it means for an input to have the correct attributes. Economic activity is subject to a great deal of **uncertainty**. The buyer's needs (and seller's costs) may change over time. For example, a pork processing firm may experience highly uncertain demand. It will need hogs delivered in the future, and the precise date of delivery will be crucial. But this date cannot be specified in advance, because the factors which determine this date are too difficult or costly to describe in a legal contract. More generally, it is impossible to specify what to do under all contingencies because human beings cannot even conceive of all the possibilities. We call this problem **bounded rationality**. These two problems, *bounded rationality and uncertainty*, mean that a contract is invariably an **incomplete contract**. Although we may want to specify certain conditions, we cannot.

3. Transaction costs and opportunism

We call the costs of writing and enforcing contracts **transaction costs**. When transaction costs are particularly high, economic actors may opt for simple arrangements which seem not to be optimal, but really are if transaction costs are taken into account along with other kinds of costs. For example, often employees do not even have written contracts, or their contracts are extremely vague compared to the verbal promises made by the employer. Rather than writing costly contracts, sometimes people must rely on such factors as reputation and the expectation of an ongoing relationship to assure that expectations are met. Unfortunately, because contracts are incomplete, people can sometimes take advantage of others, by adhering to the letter of agreements while violating the spirit. For example, a farm laborer may travel a long distance after being promised a job by a farmer hundreds of miles away. When he arrives, he may find that the job lasts only a month (when a long-term position was promised), or even if the farmer gave a written assurance of a long-term job, he may treat the worker so badly that the worker chooses to quit. We call this problem **opportunism**. In this case, it is made possible because transaction costs are too high for the farmer to define "good treatment" adequately in an employment contract. Opportunism leads to the hold-up problem, discussed below.

4. The hold-up problem

Consider again the pork processing firm. Suppose that this packer wants to contract with a number of hog farmers to deliver hogs when they are needed. To reduce transportation costs, the packer decides to situate a processing plant near the hog farms, at a significant investment. Now suppose that the packer and a farmer enter a market relationship. The contract between them cannot specify the date at which hogs must be delivered. When the packer needs a delivery, the farmer may insist on an especially high price before delivery occurs, claiming that the contract does not require her to deliver hogs at that time. The processor will be willing to pay this high price, since he has already invested in the meat processing plant and transportation costs are too high to purchase from another farm. The cost of building the processing plant are sunk. In other words, the farmer can **hold up** the packer. This problem may be so severe that the packer decides not to build the processing plant at all. (Similarly, in the case of the farm worker described above, the worker may choose not to travel to farm, even though his labor is badly needed.)

5. Asset specificity

We call the packer's investment in the processing plant a **relationship-specific investment**. (The worker's costs of moving are also a specific investment.) A relationship-specific investment is one which is made to support a given transaction. The asset purchased is then called a **relationship-specific asset**. Only when **asset specificity** exists, are incomplete contracts a source of trouble. If the packer could easily pick up his plant and move it somewhere else, the farmer would have no power, and no hold-up problem would exist.

There are many other examples of asset specificity: If a firm purchases or designs machinery to meet the terms of a specific contract, then this machinery (asset) is relationship-specific; it has much less value outside of a particular economic relationship. Consider a fruit harvesting firm that purchases specialized boxes to meet the needs of a particular packer with which it has a delivery contract. If these boxes do not work well in the operations of other packers, then this asset is relationship-specific. Or consider a data processing firm which develops software to meet the needs of a particular client. This software is then relationship-specific.

6. Solving the hold-up problem: Why vertical integration is sometimes essential.

One faulty solution to this hold-up problem would be to allow the packer to tell the farmer exactly when to deliver the hogs, with a price specified ahead of time. The problem with this is that then the packer could hold up the farmer, by insisting on delivery at very inconvenient times unless a bribe is paid. The farmer would then not be willing to invest in expanding her herd.

The solution: The packer should buy the hog farm, and hire the farmer as an employee. This way, the packer can tell the farmer when she must deliver hogs, but he will not be unreasonable since the profits of the farmer are his own profits. *Vertical integration solves the hold-up problem* in this case (but not in all cases).

7. Information: Another reason for contract incompleteness.

a. proprietary information

A firm may possess *private information* that is known to it and to no other participants in the market. Such information may be important for the firm's competitive position. KFC has long used its *secret recipe* as an effective promotion tool. Some oil additives are created with processes that the companies view as proprietary. Textile manufacturers may have specific combinations of dyes that produce their trademark colors. If an outsider is hired to use this information in production, the original firm must use a contract which forbids spread of such information. Such contracts are very difficult to enforce. The more firms that are involved in the production process, the more likely this private information is to become public.

b. asymmetric information

If one party to a transaction has information which other parties do not have, then we say that there is **asymmetric information**. For example, a purebred swine producer may have superior information about the maternal characteristics of a particular line, such as milk production, general care of young, cleanliness in the farrowing crate, etc. In a given transaction, this producer may have incentives to reveal or hide this information. It may be impossible to share this information in a way such that the buyer knows the information is accurate. This information therefore cannot be used to define the contract terms between the buyer and the seller. If it were, the contract would be unenforceable. For example, suppose that a contract specified that, if swine do not take good care of their young, the seller must refund the buyer a certain fraction of the purchasing price of the feeder pigs. After the purchase, the buyer may then claim that the swine are bad care-givers, but how can she persuade a court that she is correct if the court cannot independently verify her claim. These characteristics of the swine are **noncontractible**.

8. Problems with vertical integration: How freedom and responsibility go hand in hand.

A market firm will typically have stronger incentives to hold down costs and innovate than a division

performing the same function within a vertically integrated firm. If a market firm fails to produce efficiently it will lose business to rivals. It is subject to the **discipline of the market**. A division within a firm may view itself as having a captive market and may not make a comparable effort. In particular, a market firm may be better motivated to reduce agency and influence costs.

To the extent that a vertically integrated firm can expose its divisions to market forces, some of these costs can be avoided. Thus we often see organizations that have “profit centers”, semi-autonomous units, transfer pricing rules that mimic the market, etc. if the firm chooses to make as opposed to buy. However, it is essentially impossible to obtain all the advantages of vertical integration and all the advantages of markets. You can’t have your cake and eat it too. (See Eccles [1985].)

For example, consider the pork processor discussed above, and suppose that he buys the hog farm and hires the farmer as an employee, as is necessary to solve the hold-up problem. Now, because the farmer does not receive any profits, he may not be sufficiently innovative in finding ways to produce high quality hogs efficiently. Unfortunately, the more that the profits from the hog farm go to the farmer, the more incentive the packer will have to abuse his power over the farmer, by asking her to deliver hogs at inconvenient and costly times. If the farmer has a higher responsibility for profits, she must also have more freedom over her decisions, but this freedom can lead to a hold-up problem. Vertical integration was chosen precisely to reduce both the freedom and the responsibility of the farmer, thus assuring timely, efficient delivery of the hogs. The cost is reduced initiative by the farmer, and this cost is unavoidable.

VI. Other Issues in the Theory of the Firm

A. A note on the horizontal boundaries of the firm

While vertical boundaries refer to the degree to which a firm undertakes more than one step in a vertical chain, horizontal boundaries refer to the number of products the firm produces and the number of vertical chains in which it participates. The reasons for producing more than one product have to do with economies in multi product production, procurement and/or marketing, also called economies of scope. An example might be the farmer who finds that the joint production of corn and soybeans leads to lower production costs than monocropping of either. Similarly, a grain cooperative may find that by handling fertilizer and seed it can raise its annual revenues. These issues are best considered after discussing the cost structures of firms in more detail and so will not be addressed here.

B. Firm organization

In addition to determining its vertical and horizontal boundaries, a firm must determine how to organize itself internally to produce and distribute its products most efficiently. Consider again the example of the camper making pancakes. If this individual decides that his particular pancakes are especially tasty and that there are profits to be made in making them for the market, then he may want to consider setting up a for-profit firm. This may start out as an evening and weekend operation in which he travels to various locations such as schools, campgrounds, state fairs, etc. and markets the pancakes out of the back of his truck. The technology may be similar, but he may add a larger propane fired grill, more mixing equipment, the sale of orange juice, etc. Over time, he may decide to open a permanent location near his home. Eventually he will need to consider hiring workers. He will also need to decide which items to buy and which to prepare internally. He may also need to consider storage facilities for the materials needed in production. He may eventually need to hire a bookkeeper and a restaurant manager if he is going to concentrate on production.

Once the number of employees and contracted services increases he will need to consider alternative ways to organize these workers. This leads to classic problems in firm organization. The owner will need to

consider who reports to whom and to what extent the operation is flat or hierarchical in organization. One way to organize is along classic unitary functional lines where all similar functions such as production, transportation and sales are lumped into working groups. An alternative in the multi-division form where the firm is organized in output groups. For the pancake maker a logical division might be the restaurant group and the traveling group. Matrix forms of organization are also possible. Each form will have different advantages and problems. A more thorough discussion of these issues is the subject of courses in **organizational behavior**.

C. Ownership and control

Once a firm becomes larger, the owner may want to consider different ways to hold the firm's assets. He may choose to take on a partner or even incorporate. Each of these will have different affects on firm efficiency and competitive behavior.

In the case of a corporation, there may be a separation of ownership of the firm and control of the operation of the firm. When control is separated from ownership, the firm's managers may not be motivated to maximize profits. They may choose other objectives such as maximizing their own income, their own leisure time, their own perks, etc. The managers are workers, and as in the case of other workers, agency issues become very important. These particular agency issues are the subject of courses in **corporate finance**.

VII. References

- Akerlof, George A., and Yellen, Janet L. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100 (supplement): 823-38.
- Baumol, W. J., J. C. Panzar, and R. D. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich Inc.
- Eccles, Robert G. 1985. *The Transfer Pricing Problem: A Theory for Practice*. Lexington, MA: Lexington Books.
- Färe, R., S. Grosskopf, and C. A. Knox Lovell. 1986. "Scale Economies and Duality." *Zeitschrift für Nationalökonomie* 46: 175-182.
- Ferguson, C. E. 1971. *The Neoclassical Theory of Production and Distribution*. Cambridge: Cambridge University Press.
- Hanoch, G. 1975. "The Elasticity of Scale and the Shape of Average Costs." *American Economic Review* 65: 429-97.
- Hoch, I. 1976. "Returns to Scale in Farming: Further Evidence." *American Journal of Agricultural Economics* 58: 745-749.
- Panzar, J. C., and R. Willig. 1977. "Economies of Scale in Multi-Output Production." *Quarterly Journal of Economics* 91: 481-493.
- _____. 1981. "Economies of Scope." *American Economic Review* 71: 268-272.
- Shephard, R. W. 1953. *Cost and Production Functions*. Princeton: Princeton University Press.
- _____. 1970. *Theory of Cost and Production Functions*. Princeton: Princeton University Press.
- _____. 1973. *Indirect Production Functions*. Meisenheim am Glan: Verlag Anton Hain.
- Weiss, Andrew. 1990. *Efficiency Wages: Models of Unemployment, Layoffs, and Wage Dispersion*. Princeton: Princeton U Press.