Health Economics

Dr Pratap C Mohanty

Department of Humanities and Social Sciences,

Indian Institute of Technology Roorkee

Week – 03

Lecture 14- Production and Costs of Health Care

Welcome, friends. Once again, regarding our NPTEL MOOC module on health economics, this is a 12-week program that is a four-credit equivalent course. You are exposed to learn various aspects of health economics. This week, we are emphasizing supply in healthcare and its perspectives or discussions. In this particular lecture, we are targeting to explain production and cost of healthcare and how students or scholars used to follow from different readings.

So, what we did in the previous lecture was on the physicians and hospitals in the healthcare market as health and healthcare providers. So, the particular target of this lecture is how economic analysis of production and cost are applied to healthcare. This will indeed contribute to the efficiency analysis, market failure discussions, economic evaluation and theory of supply and market structure. What is all about the production and cost? We will start with the basic principles of production or the production theory.

In the start, we wanted to just mention that economic theories of supply can also be extended to healthcare as an economic good. Healthcare comprises an extremely diverse range of goods and services, including screening programs, surgical procedures, pharmaceuticals, and counseling services, and we refer all these to as products. There are many different types of healthcare providers, including hospitals, GP practices, retail companies and ambulance services. We refer to those subparts of healthcare as firms as like our industry. So, we are addressing different types of healthcare with each category of healthcare.

The quantity of healthcare product produced by healthcare firm is referred to as its output. So, to define healthcare output as part of healthcare facilities, though it is a very tedious task, however, to simplify this, there are two approaches followed. One is called change in the level of health and amount of care provided and usually considered as the intermediate output. So, how production functions are defined? What does production function do? It exhibits the relationship between inputs that goes into the manufacturing process and generate output and in this context we are identifying specific outputs of healthcare industry. So, what is an input? In this case, used to be the factors in production function and

usually, they are called labor, land, raw materials, and capital within a Hospital, what are inputs in the context of the Hospital we are mentioning?

What are the inputs in producing cataract removals in a surgical unit? So, it is the labor involved for such certain labor hours, then anesthetist and nurses, and there will be capital involved as well, like types of machinery, hospital beds, etc., instruments used at the theatre, and operating theatre. There will be materials also involved, such as drugs or medicines, dressings, and disposables, and these all constitute the inputs that these inputs produce and what is the output here in this case? Basically, the simple target in this example is better eye health. Can we measure eye health perfectly? What if the operation is unsuccessful, and in that case, what is the output? So, the intermediate output is the number of operations we did. Generally, these are the quantity figures required when we operate efficiency analysis or to find out efficiency or productivity analysis. So, quantity figures are usually required. And we know that there are a number of directions where quantitative estimations are difficult.

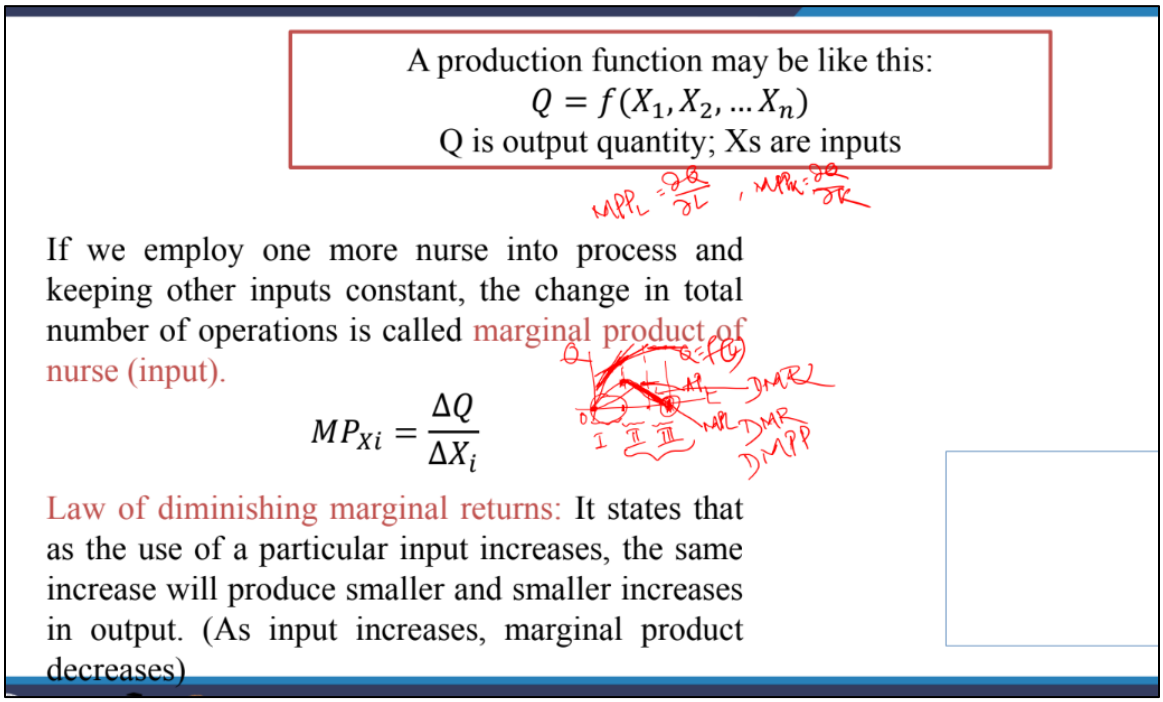
We need to go by some qualitative approximations, maybe by developing indexes or taking proxy variables. When we study the production process in healthcare, several questions arise. So, to what extent can different inputs such as surgeons and nurses, be substituted? Now, we are leading the discussion in terms of their input combinations of the substitutions to have a better performance indicator. So, if we employ more nurses, how many more operations will be able to perform? How many more operations will be able to perform if we use more of all of the inputs? How can we tell if production is being carried out efficiently? This is what I just mentioned. Hence, the production function can be presented in an equation like Q as a function of X_1, X_2 and X_n.

Q is your output in quantity, and X's are input. X may be, as I already told you, the labour hours of the surgeon or the nursing staff, or maybe X_2 or another capital use, machines, or operating theatre itself. So, similarly, other factors may be that building is required, and beds are required. Finally, output, as I already mentioned in that case, we are taking the incidence or the number of cases treated. So, if you employ one more nurse into the process and keep other inputs constant, the total number of operation changes is called the marginal product of the nurse or input.

As we already read, in simple microeconomic theory, you must have read marginal physical productivity of labor, which is defined as $\frac{dQ}{dL}$. You might have also read MPK with respect to $\frac{dQ}{dK}$. So, we are just using a similar concept here just to clarify. The marginal productivity of the i-th inputs is nothing but $\frac{\partial Q}{\partial X}$, X is our input. In an obvious context where the natural laws of production follow the principles of diminishing marginal returns, there are different stages of the production process. If you remember, there are first, second, and third stages in the production function.

I am not making you more confused. I think when your L we used to take and Q we used to take here, when L is 0, we used to get 0 as the output. So, we used to draw this type of production function and reached a maximum point Q. We used to say Q is a function of L (Labor), and K (Capital) is what we say as fixed and then normal production function. If you remember a microeconomic theory, there are different inflection points that precisely identify the changes from one stage to another. There are some stages that identify the maximum production.

We refer to dQ by dX or $(\frac{dQ}{dX})$ equal to 0, and if you remember, we used to draw this type of diagram to identify marginal productivity of labor, etc. We used to draw the average productivity of labor (APL) and accordingly, we derive the stages of production, the inflection I have just derived mentioned here. Inflection point should occur at this point since the maximization reaches here and somewhere, and another maximization occurs due to the average productivity. So, another inflection point can be presented and the third one is here. So, accordingly we derive first phase, second stage and third stage.



Stage 1, stage 2 and stage 3. I am just recapping the microeconomic theory. But here somewhere we are touching up this concept and to differentiate how healthcare is little different, but largely it follows the production function. The diminishing marginal product is the obvious outcome of the production function because in the first stage we generally do not target to be our objective function because others if we just target and try to maximize the first stage we used to be underutilizing our capacity. So, we need to give the space for capacity extension.

We try to reach at the maximum possible point. Hence, the diminishing phase is the most important effective hours for determining the best combination of labor for the output. Hence, the production function is exhibited during the DMU phase. But in this case, DMR is the diminishing marginal returns phase. So, this states that as we keep increasing the input, the output will change in different proportions. Since it gets declining rate in the output after a certain time, that follows the law of diminishing marginal return phase.

This is also called diminishing marginal returns and diminishing marginal productivity since we are emphasizing productivity. In some book you will find diminishing marginal physical productivity of L etc. These kinds of words you might find. This is precisely the figures taken from the work of the World Bank 2009. It is based on total health expenditure and output, which is life expectancy, and the relationship is that health expenditure has diminishing returns.

So, just for you to check and if you remember we also explained we used to understand in the simple microeconomic book. You can still easily cover it even if you have not read it. Isoquant simply presents the combination of factors, factor input that result based on output. So, the combination is basically the mapping of the combination of factors that result into the optimum level of output and or indirectly, once the optimum output is defined, what are the best combinations possible that will result into the same quantity every time. In this case, we have taken two factors.

They are doctors and nurses. For each combination of this input, some output or operations will be performed. Isoquant for different combination of two inputs if output is the same and that is termed as Isoquant. Basically, output is the same. In the Isoquant, the output has to be same and we are trying to define the best combinations and it is presented here.

This curve maps the output for different combination of nurses and doctors. And this is called Isoquant. Iso stands for equal and Q stands for quantities. All combinations are deriving equal quantities. So, there are possibility of substitutability between these two inputs and slope at any point on Isoquant measures the degree to which different factors of production can be substituted for each other.

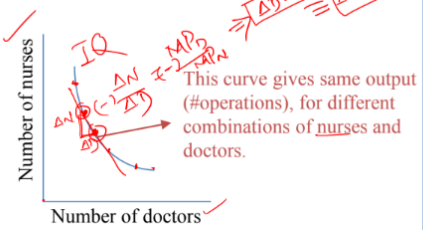
So, the slope is basically, if you take a slope anywhere, that is precisely this, let it be and this slope of Isoquant is called marginal rate of technical substitutions. So MRTS, marginal rate, so basically from the marginal changes can be delta, doctors here, delta nurses. So, here basically it is $\frac{\Delta N}{\Delta D}$ (N and D indicates nurses and doctors, respectively *(fig drawn by professor in slide)*). So, over here *(as given in slide notation, pointing out in the formula)* Δj, j is our nurses and i stands for doctors *(i is on horizontal axis and j on vertical axis)*. This is precisely for the best combinations to be attended.

Isoquants

Consider a case in which a hospital has only two factors of production: Nurses and Doctors

For each combination of these inputs, there will be some output or operations performed.

Isoquant: For different combinations of two inputs if output is same, it is termed as an isoquant.



This curve gives same output (#operations), for different combinations of nurses and doctors.

Substitutability between inputs

Slope at any point on isoquant measures the degree to which different factors of production can be substituted for each other.

- Slope of isoquant is called marginal rate of technical substitution (MRTS)

$$MRTS_{XiXj} = \frac{\Delta X_j}{\Delta X_i} = -MP_{X_i}/MP_{X_j}$$

- Isoquant is convex to origin
- MRTS decreases as we move from left to right along the isoquant.

The change in the technical substitutions must correspond with the change in their marginal productivity. So, the marginal productivity has to be equalized. This has to be equalized with the marginal productivity of D with respect to marginal productivity of N. Why are we just making a reverse and the slope used to be negative and this is also negative. And in that case, we used to derive total productivity like delta D times marginal productivity of D should be equalized with delta N times marginal productivity of N.

Because the change in from one point to another point, the total output remains constant, hence these two sides should be equalized. But hence, because this is nothing but the total Q, the Q generated out of the change. So delta Q, so delta Q has to be same. Hence it is nothing but the ratio. So ratio is simply highlighted in this figure.

So isoquant is convex to the origin and to the origin, we are actually mapping to the origin and there are even properties why it is convex, you have to take first order and second order derivative and understand the changes. MRTs decreases as we move from left to right along with the isoquant. So the returns to scale of production, we discussed first the basic isoquant. Then there are scale of returns or returns to scale of production exhibit differently in different stages. We used to read a concept called constant returns to scale, then increasing returns to scale and decreasing returns to scale.

So, constant basically follows the property where the production function holds with CRS when a proportional increase in all inputs result in an increase in output by the same proportion. In an example, we have said that 10 times increase in nurses or nurses and doctors is a combination or other inputs will result in 10 times as many as operation performed. That means the output change by 10 times since we change the inputs by 10 times. That is basically the answer for CRS. Whereas increasing returns to scale is basically your output increases more than proportionately than that of your change in proportion of

your inputs.

So if operations basically our output is Q, if operation increase is by 11 percent then that of your input is of 10 percent we follow the IRS, increasing returns to scale function. In decrease one if it is just the reverse. This is what is explained in this paragraph. I think I have already explained. However, when we see increasing returns to scale, two main factors account for increasing returns to scale.

One is called indivisibility and specialization. So, indivisibility we mean because of its technical and managerial indivisibility. This means that there is a certain minimum size of the factor, and even if it is large in relation to the output size, it must be used. Like a body scanner in a rural facility where there are only a few patients may not be used to its technically possible capacity. Body scanner are lumpy in nature and usually it is technically indivisible, hence it requires more number of patients to be scanned.

But if you are just scanning very few then its efficiency is not observed. Hence the efficiency, the better use and the value derived when or even the efficacy of it is derived when you apply for more patients. Usually specialization due to division of labor is related to divisibility issues. Usually the variable inputs are considered to be highly divisible. However, in the long run, they are also considered divisible in the fixed factor.

So, what is the point of explanation in this case is that your factors are otherwise known for their contribution in specific field. Some of the factors can be allotted or allocated with a specific domain and they will used to perform very well. Like a small hospital may have to employ general surgeons to deal with many areas of surgery, but a larger one may be able to employ a specialist surgeon instead of just as normal surgeon who become proficient in performing specific type of surgery and can produce better output and there might be more consultation because of name and fame. And constant we have already discussed, but economists believe that if there is less inefficiencies of production, no indivisibility problem, no specialization issue, then expansion in scale leads to a situation where returns to scale increase in constant returns to scale pattern. And in the DRS case, we already said the output changes are comparatively lesser than that of the change in the inputs.

It might be due to your managerial difficulties or due to the problem at the enterprise level. So, managerial basically when the scale of production expands, the coordination and the control on different factors of production tend to become weak and output fails to increase. In the same proportion as the factors of production increase. In the enterprise context, this region considers enterprise as a whole an individual factor of production. When other factors increase, enterprise becomes relatively scarce and leads to decreasing returns to scale function.

We have discussed about its Q or the output, its production function. Let us also explain to some extent about what is called cost. Economics measure the value of resources by their

cost. If for an input unit cost is P and Q units are used in the production, then the total cost will be of course P times Q. Hence, the total cost is presented with the notation as C will be of course your first inputs cost and its input that is X1 and its price then X2 its respective price and so on till the nth inputs and the price already I have mentioned.

If there are two inputs that those are nurses and doctors in our case, then cost will be of course price of Q and nurses, price of each per unit of nurses are hired and the doctors are taken their price, then we can draw the respective cost line and it is linear so we can easily present in a linear diagram and everywhere on this particular line, this is nothing but the equation which I have just mentioned and this is called isocost because in all the points are explaining the same level of cost. And the slope of the isocost line is nothing but the minus Pd by Pn. So you can find out some clarity, you can just take the first order derivative because this is nothing but you can just take the tan theta of it, this is P, sorry tan theta which is nothing but P perpendicular times P by Q. So here it is your P perpendicular, Q is the horizontal distance this by this. So this area you can find out by taking this distance.

If you put it the values in the cost function and you can easily find out the distance. And so basically this is the maximum possible quantity of nurses if doctors are 0, if you make it Q_d to be 0 then the maximum difference can be derived and so on, slope can be easily derived. So, cost function which I have already said and some of the important concept we wish to get it for understanding the cost function that is called average cost, AC that is total cost divided by number of inputs and used to be deriving through individual average cost maybe with respect to variable cost, AC L (Labour) or AC capital (K). And marginal cost then it will be incremental change then delta to be used. So you can also follow some of the microeconomics book to answer through some example is to give tables to answer.

Here given different cost combinations or cost labels our purpose is to minimize the cost. So in this case the third one, the uppermost one is not required to be attended since our target output can be reached at the respective cost function. The choice of factors that minimize the production cost can be determined by the point on the isoquant that is the lowest associated isocost. At that particular point this slope has to be maintained. The slope of the isoquant has to be equalized with the slope of the isocost line.

We have already seen the isocost line and slope is P_d by P_n that should be equalized with the slope of the isoquant that is marginal probability of d divided by marginal probability of n. So, we seek to minimize total cost given a fixed product output. So, our objective function suppose some mathematical questions you can try. If your objective function is to minimize the cost then it should be the C function is objective function. Then C is nothing but we already said the minimization of this is nothing but the C price times quantity of nurses plus price of doctors times quantity of doctor given a level of output given a fixed output that is Q_0 let it be at a level Q_0.

We can follow Lagrangian since the optimization function if it is not linear. In that case

sometimes C comes with some of the equations such as I am just writing here C might be presented as 100 plus so maybe 2 times C_1 or X_1 maybe or 3 times X_2 square or maybe anything like this equation is given and this seems to be a non-linear equation. Non-linear if your objective function is non-linear with a constant linear function, but not in a unconstrained linear function. We are referring to a constant linear function that is our production function assumes a level. So, in that case, but here we have taken what we have taken, but for your simplicity we have taken a linear cost function and objective function in non-linear.

So in that case one has to be linear and there has to be non-linear in that way we can able to optimize the point and find out through a Lagrangian multiplier function. But here what we did it as per the Lagrangian formula, we followed first our linear objective function in this case for simplicity we have taken the linear one because our target is to minimize the cost plus the Lagrangian multiplier has to be given and this is nothing but equating to the production function. And we need to fulfill this dL by Lagrangian dL by dQ with respect to n, with respect to d and with respect to lambda. All has to be equated to 0 and we can find out and solve it and find out the solution and this will precisely give you the optimization function we just mentioned here. You can try using the right example, and these are also available in microeconomics books.

Cost minimization: Mathematical analysis

We seek to minimize total costs given a fixed product output:

$$\min P_N Q_N + P_D Q_D = C$$

Given $f(Q_N, Q_D) = Q_0$

Setting up the Lagrangian

$$L = P_N Q_N + P_D Q_D + \lambda(Q_0 - f(Q_N, Q_D))$$

$$\frac{\partial L}{\partial Q_N} = \frac{\partial L}{\partial Q_D} = \frac{\partial L}{\partial \lambda} = 0$$

Solving this

$$\frac{P_D}{P_N} = \frac{MP_D}{MP_N}$$

If in the previous problem production function follows cobb-douglas one, if your production function is this, find the equilibrium condition of cost minimization. So, this is what we have just mentioned. This function if it is cobb-douglas one then you can apply. Like your constraint function is given with this, a cobb-douglas production function then the same approach by Lagrangian multiplier. So, start with the cost function then lambda times, if you follow the Lagrangian multiplier formula you will find out.

So, what are the factors affecting the cost of hospital services mentioned by Fournier and

Mitchell in 1992, estimated multiproduct cost function for 179 short term general care hospitals in Florida from the year 1984 to 1986. They found more output, it requires more cost, inpatient admission, outpatient visits, then emergency room visits, surgery, minutes etc have positive effect on cost. Similarly other factors you can just read what are the findings. I am not going by the points.

They also included about effect of competition. I think those will be useful for you to read and understand. So, to follow it very correctly, those are novice readers and just to start reading microeconomics and health economics in particular you have to read these carefully. In the next lecture we will discuss about profit maximization model in healthcare market. With this I think I must stop here. Thank you.