Beyond homogeneous production functions

Gluschenko, Konstantin

Institute of Economics and Industrial Engineering, Novosibirsk State University

5 November 2025

Beyond homogeneous production functions

Konstantin Gluschenko

Institute of Economics and Industrial Engineering;
Novosibirsk State University
Novosibirsk, Russia
E-mail: glu@nsu.ru
ORCID: 0000-0003-1209-3809

Abstract

Standard production functions, being homogeneous, have unchanging returns to scale. Therefore, it remains unclear to students where the U-shape average-cost curves come from. This article considers non-homogeneous production functions that have variable returns to scale and can produce such curves, providing a specific example. The article also discusses the link between economies/diseconomies of scale and market structures, showing that economies of scale are incompatible with perfect competition. They are peculiar to monopolies and oligopolies, whereas in a perfectly competitive market there are no economies of scale, diseconomies of scale being a quite normal phenomenon.

Keywords: returns to scale; economies/diseconomies of scale; non-homogeneous production function; firm's expansion path; market structure

JEL codes: D24; D40

Introduction

All microeconomics textbooks discuss the long-run average-cost curve, showing that it has a U-shape. Thus, as output expands, returns to scale change from increasing to constant (maybe, at a single point) and then to decreasing. At the same time, the standard production functions under study, being homogeneous, have unchanging returns to scale. Therefore, it remains unclear to students where the U-shaped average-cost curves come from.

Furthermore, there appears to be an “asymmetry” in the consideration of changes in average costs with increasing output, namely, too much attention is paid to economies of scale, underexposing diseconomies of scale. However, it is the latter (as well as the absence of economies/diseconomies) that is more important, as it is a necessary condition for a firm to operate in a perfectly competitive market. I would note that the lack of a general term characterizing the dependence of average costs on scale of production is somewhat inconvenient. Such a term could be, for example, “the average-cost effect of scale.” Depending on the sign of the rate with which average costs change, the effect may be called negative when they decrease (economies of scale), zero when they remain unchanged, and positive when they increase (diseconomies of scale).

In this article, I discuss non-homogeneous production functions in which returns to scale are variable and which, therefore, can produce U-shape average-cost curves. An example of such a curve produced by a specific function is given. The link between returns to scale and market structure is considered graphically. This shows that economies of scale occur in the case of a monopoly or oligopoly, whereas perfect competition is characterized by the absence of economies of scale.

Throughout the article, a long-term period is implied, so average costs

everywhere are average total costs. To avoid complicated mathematics, two-input production functions are dealt with. The input prices are assumed to be constant as output expands, which ensures a direct connection between returns to scale and economies/diseconomies of scale. (The case of variable input prices received attention in Cohn 1992, Truett and Truett 1995, Gelles and Mitchell 1996, and Comolli 2000.)

The article is organized as follows. The following section presents basic relationships characterizing returns to scale and economies/diseconomies of scale. Next, I consider non-homogeneous production functions that can produce U-shape average-cost curves. Further, I discuss the link between economies/diseconomies of scale and market structures. Finally, I provide conclusions.

Preliminaries

Let $Q(L, K)$ stands for a production function, L being labor and K being capital. The notation $\varepsilon[y, x]$ will mean the elasticity of y with respect to x. Denoting the scale of production in terms of inputs with s, the relationship $Q(sL, sK) = s^{\varepsilon}Q(L, K)$ gives an idea of the returns to scale: with $\varepsilon < 1$, $\varepsilon = 1$, and $\varepsilon > 1$, we have decreasing, constant, and increasing returns to scale, respectively. It is easy to see that ε is the scale elasticity of output: $\frac{dQ(sL,sK)}{ds} = \frac{d(s^{\varepsilon}Q(L,K))}{ds} = \varepsilon s^{\varepsilon-1}Q(L,K) = \varepsilon\frac{s^{\varepsilon}}{s}Q(L,K) = \frac{\varepsilon}{s}Q(sL,sK)$, hence, $\varepsilon = \frac{dQ(sL,sK)}{ds}\frac{s}{Q(sL,sK)} = \varepsilon[Q,s]$.

The scale elasticity of output can be represented in other forms that are more convenient for calculation, not requiring explicit involvement of the scale factor. First, the scale elasticity of output equals the sum of elasticities of output with respect to labor and capital, $\varepsilon[Q, s] = \varepsilon[Q, L] + \varepsilon[Q, K]$ (this relationship is sometimes referred to as the Wicksell-Johnson theorem): $\varepsilon[Q,s] = \frac{dQ(sL,sK)}{ds}\frac{s}{Q(sL,sK)} = \left(\frac{\partial Q}{\partial(sL)}\frac{d(sL)}{ds} + \frac{\partial Q}{\partial(sK)}\frac{d(sK)}{ds}\right)\frac{s}{Q} =$

$\left(\frac{1}{s}\frac{\partial Q}{\partial L}L + \frac{1}{s}\frac{\partial Q}{\partial K}K\right)\frac{s}{Q} = \frac{\partial Q}{\partial L}\frac{L}{Q} + \frac{\partial Q}{\partial K}\frac{K}{Q}$. Second, it can be also shown that the scale elasticity of output equals elasticity of output with respect to total costs, $\varepsilon[Q, s] = \varepsilon[Q, TC]$.

Consider average costs, assuming for simplicity that the cost-minimizing ratio of inputs is constant: $AC(sL, sK) = \frac{TC(sL,sK)}{Q(sL,sK)} = \frac{sTC(L,K)}{s^{\varepsilon[Q,s]}Q(L,K)} = s^{1-\varepsilon[Q,s]}AC(L, K)$. Therefore, the scale elasticity of average costs is $\varepsilon[AC, s] = 1 - \varepsilon[Q, s] = 1 - \varepsilon[Q, TC]$. The relationship derived by Gelles and Mitchell (1996) links the elasticity of average costs with respect to output with the scale elasticity of output: $\varepsilon[AC, Q] = \frac{1-\varepsilon[Q,s]}{\varepsilon[Q,s]}$. On the other hand, $\varepsilon[AC, Q] = \varepsilon[TC, Q] - 1$ (since, obviously, $\varepsilon[TC,Q] = 1/\varepsilon[Q, TC]$). All these formulas suggest the well-known fact that increasing returns to scale ($\varepsilon[Q, s] > 1$) correspond to economies of scale ($\varepsilon[AC, s]$, $\varepsilon[AC, Q] < 0$), decreasing returns to scale ($\varepsilon[Q, s] < 1$) correspond to diseconomies of scale ($\varepsilon[AC, s]$, $\varepsilon[AC, Q] > 0$), and constant returns to scale ($\varepsilon[Q, s] = 1$) correspond to the absence of economies and diseconomies of scale ($\varepsilon[AC, s]$, $\varepsilon[AC, Q] = 0$). Recall that this is true only if the input prices remain constant.

Variable returns to scale

Scale elasticities of output for standard production functions, such as the Cobb-Douglas, Leontief, and CES functions, are constant (hence, returns to scale are invariable). This implies that these functions are homogeneous according to the definition of such functions: $f(sx_1,\ldots, sx_n) = s^{\varepsilon} f(x_1,\ldots, x_n)$ with constant ε. Homogeneous functions are always homothetic. This entails that the firm's expansion path is a straight line passing through the origin: $K = cL$ or $K/L = c$, where c is a constant that depends on input prices and parameters of a given production function (for instance, $c = \frac{w}{r}\frac{\beta}{\alpha}$ for $Q = L^{\alpha}K^{\beta}$, w

and r standing for prices for labor and capital). Just this lets all inputs grow at the same rate to ensure minimal total costs.

The constancy of $\varepsilon[Q, s]$ leads to the constancy of $\varepsilon[AC, s]$. To describe variable returns to scale, non-homogeneous (with varying ε) production functions are needed. These are rarely mentioned even in advanced microeconomics textbooks (one of the exceptions is Jehley and Reny, 2011), although only such functions can produce U-shape average-cost curves that are shown in all microeconomics textbooks. The relationships $Q(sL, sK) = s^{\varepsilon[Q,s]}Q(L, K)$, $AC(sL, sK) = s^{1-\varepsilon[Q,s]}AC(L, K)$, and other relationships from the previous section still hold for non-homogeneous production functions, however locally. Their local versions can be defined strictly. For instance, such definition of the scale elasticity of output at the point (L, K) is $\varepsilon[Q, s] = \lim_{s\to 1} \frac{d\ln(Q(sL,sK))}{d\ln(s)}$.

Figure 1 provides examples of two non-homogeneous production functions. Such functions usually are non-homothetic; the expansion path is then either nonlinear as in Figure 1(a) or linear but not passing through the origin as in Figure 1(b). However, this is not always the case. An example of non-homogeneous but homothetic function is $Q = \ln(L^{\alpha} + K^{\alpha})$ with the expansion path $K/L = (r/w)^{\frac{1}{\alpha-1}}$.

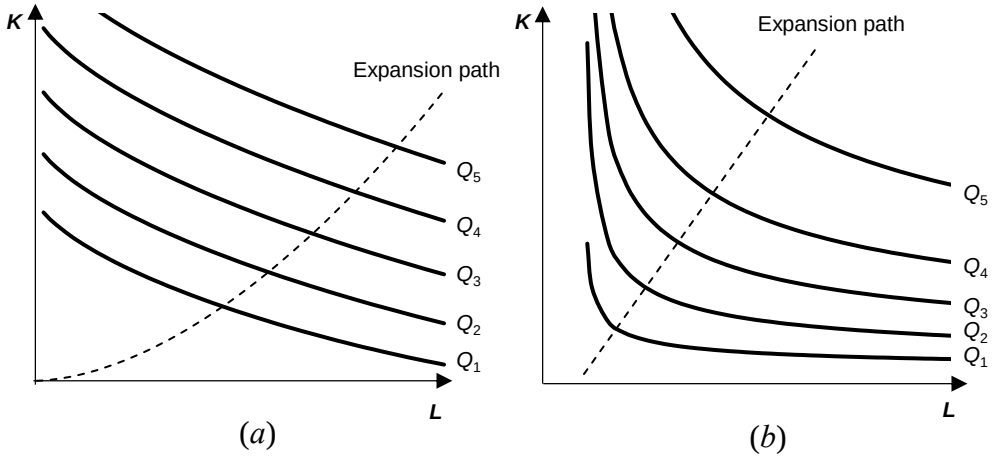

Figure 1. Isoquants of non-homogeneous production functions:
(*a*) $Q = L^{\alpha} + K^{\beta}$, (*b*) $Q = (L - L_0)^{\alpha}(K - K_0)^{\beta}$.

Let us consider function $Q(K, L) = (L - L_{\min})^{\alpha} (K - K_{\min})^{\beta}$. The cost-minimization condition $\frac{\partial Q/\partial L}{\partial Q/\partial K} = \frac{\alpha}{\beta} \cdot \frac{K-K_0}{L-L_0} = \frac{w}{r}$ yields the expansion path $K = \frac{w}{r}\frac{\beta}{\alpha}(L - L_0) + K_0$. Benefiting from the relationship $\varepsilon[Q, s] = \varepsilon[Q, L] + \varepsilon[Q, K]$, we get the following expression for the scale elasticity of output: $\varepsilon[Q, s] = \frac{\alpha}{1-L_0/L} + \frac{\beta}{1-K_0/K}$ (L and K are always assumed to lie on the expansion path). With specific parameters of the function, by changing L, we can compute $Q(L, K(L))$ and $AC(Q(L, K(L)))$ as well as the elasticities (for instance, with the use of Excel). Figure 2(a) depicts a graph of $AC(Q)$ for $Q = (L - 25)^{0.25}(K - 25)^{0.55}$ with $w = 70$ and $r = 25$. Figure 2(b) shows two respective elasticities.

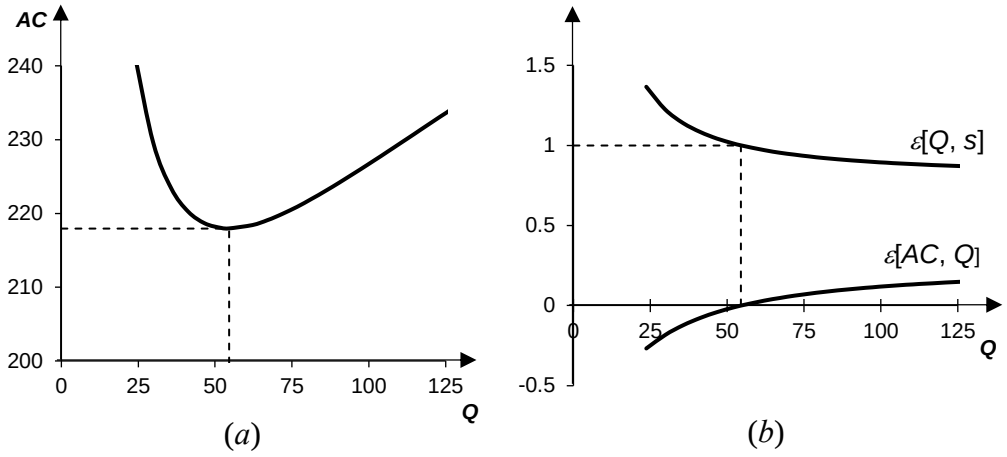

Figure 2. Properties of production function $Q = (L - 25)^{0.25}(K - 25)^{0.55}$ depending on output: (*a*) average costs, (*b*) scale elasticity of output and elasticity of average costs with respect to output.

In this example, the slope of the average-costs curve is downward when $Q < 54.5$ and upward when $Q > 54.5$. The point $Q = 54.5$ separates the regions of increasing and decreasing returns to scale. As Q increases, the scale elasticity of output decreases from values greater than 1, reaching 1 at that point (where the returns to scale are constant) and then decreasing further. The elasticity of average costs with respect to output rises from negative values, reaching zero at $Q = 54.5$ and then becoming positive.

Returns to scale and market structure

As mentioned in the introduction, economies of scale are often overemphasized in microeconomics teaching. This can give students the impression that economies of scale are always a positive phenomenon, while diseconomies of scale are a negative one. The latter is even explicitly stated in some textbooks. For example, “Diseconomies of scale can arise because of coordination problems that often occur in large organizations” (Mankiw, 2024, p. 282). This misconception can be avoided by linking economies of scale to market structure.

Figure 3 shows a U-shape average-cost curve with a flat segment, similar to that presented in Mankiw (2024, p. 280). Obviously, only a non-homogeneous production function can yield such curve. Since, in general, an increase in inputs in the same proportion is not optimal for such functions, it is better to characterize the scale of production by total costs rather than by volumes of inputs.

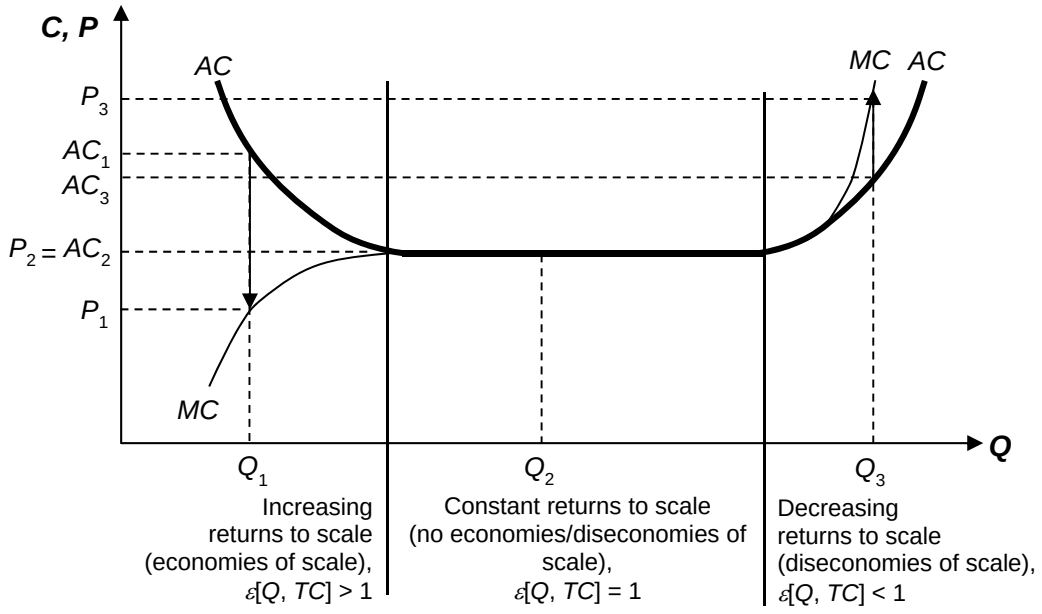

Figure 3. Average and marginal costs vs returns to scale.

In addition to the average costs, this figure contains a curve of marginal costs,

MC. A certain combination of *AC* and *MC* at a given output clearly shows which market structure it corresponds to. As is well known, $AC > MC$ when average costs decrease, $AC < MC$ when they increase, and $AC = MC$ when *AC* are at a minimum. (The value of $\varepsilon[Q, TC]$ represents these relationships, as it equals AC/MC.)

Since the profit is $\pi = (P - AC)Q$, for a firm to be profitable or at least break even, it is necessary that $P \geq AC$. On the other hand, $P = MC$ in perfectly competitive markets. However, this leads to the firm's unprofitability when its output is in the region of increasing returns to scale. For, say, $Q = Q_1$, we have $P_1 - AC_1 = MC(Q_1) - AC_1 < 0$. Hence, *economies of scale are incompatible with perfect competition*. For $P \geq AC > MC$, the firm must be a monopoly or an oligopolist.

Perfect competition is possible if firm's output is in the region of constant or decreasing returns to scale. In the first case, the firm does not gain economic profit. Provided that $Q = Q_2$, $P_2 - AC_2 = MC(Q_2) - AC_2 = 0$. In the second case, the economic profit is positive: $P_3 - AC_3 = MC(Q_3) - AC_3 > 0$. Thus, diseconomies of scale are a quite normal situation, not caused by any problems in firm's activities. It may turn out that it is profitable for the firm to increase output from some $Q < Q_3$ to Q_3 in order to satisfy the profit-maximization condition $MC = P$, even though the average costs will increase.

Conclusions

To obtain U-shape average-cost curves, non-homogeneous production functions are needed. Their scale elasticities of output are variable; nevertheless, the relationships that hold for invariable returns to scale are locally valid. Since increasing inputs in such functions in equal proportions does not, as a rule, yield a firm's expansion path, it is more convenient to characterize the scale of production not by the amount of factors, but by total costs.

To demonstrate the behavior of non-homogeneous production functions, a simple generalization of the Cobb-Duglas function $Q = (L - L_0)^{\alpha} (K - K_0)^{\beta}$ with $\alpha + \beta < 1$ may be used. It has a transparent intuitive interpretation: L_0 and K_0 are the amounts of labor and capital necessary to start production, but which do not themselves produce a product (for example, L_0 can be a minimal team of engineers and managers, and K_0 can be a factory building).

Regions of increasing returns to scale (economies of scale) and non-increasing returns to scale (the absence of economies/diseconomies of scale and diseconomies of scale) correspond to different market structures. Perfect competition is impossible with economies of scale; in this case monopoly or oligopoly takes place. Diseconomies of scale and constant average costs are more common, being typical for firms in perfectly competitive markets.

References

Cohn, E. 1992. Returns to scale and economies of scale revisited. *Journal of Economic Education* 23 (2): 123-24.

Comolli, P. M. 2000. Pecuniary effects, second-order conditions, and the LRAC Curve. *Journal of Economic Education* 31 (2): 131–43.

Gelles, G. M., and D. W. Mitchell. 1996. Returns to scale and economies of scale: Further observations. *Journal of Economic Education* 27 (3): 259–61.

Jehley, G. A., and P. G. Reny, 2011. *Advanced microeconomic Theory*. 3rd ed. Harlow, UK: Pearson.

Mankiw, G.N. 2024. *Principles of economics*. 10th ed. Boston, MA: Cengage Learning, Inc.

Truett, L. J., and B. D. Truett. 1995. Exploring the relation between returns to scale and pecuniary economies/diseconomies. *Journal of Economic Education* 26 (3): 140–49.