

NAKE Course Banking & Finance

Lecture 1: Bank Productivity and Efficiency

Jaap Bos (j.bos@maastrichtuniversity, 043-3883838)

Utrecht, October 28, 2011



Basic Model of Bank Performance

Basic Model of Bank Performance (cont.)

- This section develops a basic model of a profit maximizing bank derived from Cowling (1976), Cowling and Waterson (1976), and Stigler (1964).
- The model by Cowling describes a relationship between industry performance and market concentration, both over time (intra-industry) and between industries (inter-industry).
- Equilibrium conditions from this model can be used to test more extreme models, namely perfect competition and myopic oligopoly behavior (the classic Cournot model).
- Without loss of generality, we assume all costs to be variable costs (in the long run), and all outputs to be perfect complements with zero cross-price elasticity.
- For now, banks are also assumed to be myopic (we will later relax this assumption).



Outline

- 1 Basic Model of Bank Performance
- 2 Production theory
 - Production sets: defining inputs and outputs
 - Transformation functions: efficiency and productivity
 - Pricing opportunity sets: technical versus allocative efficiency
 - Frontier shifts: (augmented) technical change
- 3 Production functions and cost functions
 - Diamond-McFadden ...
 - Cost functions: applying duality
 - Profit or revenue functions: pricing power
- 4 Efficiency of Banks
 - X-Efficiency
 - Scale and Scope Economies
- 5 Synthesis

Basic Model of Bank Performance (cont.)

Bank i then maximizes:

$$\begin{aligned} \Pi_i &= pY_i - w_iX_i, \text{ subject to} \\ T(X_i, Y_i) &= 0 \\ H(p, Y_i, w_i, Z_i) &= 0 \\ p &= f\left(\sum_{i=1}^N Y_i\right) = f(Y) \end{aligned}$$

where $f(Y)$ is inverse market demand and N the number of banks.



Basic Model of Bank Performance (cont.)

Profits are maximized if:¹

$$\frac{d\Pi_i}{dY_i} = p^* + Y_i f'(Y) \frac{dY}{dY_i} - w_i \frac{dX_i^*}{dY_i} = 0 \quad (2)$$

Where the optimal number of inputs X_i^* depends on the demand for outputs Y_i . Multiplying by Y_i yields:

$$p^*Y_i - w_i \frac{dX_i^*}{dY_i} Y_i = -(Y_i)^2 f'(Y) \left(\frac{dY}{dY_i}\right) \quad (3)$$

where revenue is denoted by pY_i .



¹Here f' denotes the first derivative of f .

Basic Model of Bank Performance (cont.)

The corresponding Lagrangian system can be written as:

$$L\Pi_i = pY_i - w_iX_i - \zeta T(\bullet) - \theta H(\bullet) \quad (1)$$

Solving for p and X simultaneously yields the optimal output prices and input quantities (denoted by asterisks):

$$\begin{aligned} p^* &= p(Y_i, w_i, Z_i) \\ X_i^* &= X_i^*(Y_i, w_i, Z_i) \end{aligned}$$



Basic Model of Bank Performance (cont.)

Here, banks are assumed to face perfectly competitive input markets, but operate in output markets where price differentiation is potentially possible.

Thus, banks may compete via their output pricing strategies, by adjusting prices and fees according to market conditions.

We start by defining λ_i as follows:

$$\frac{dY}{dY_i} = 1 + \frac{d \sum_{j \neq i} Y_j}{dY_i} = 1 + \lambda_i \quad (4)$$

where λ_i is known as the conjectural variation of firm i 's output.

A high λ_i means a firm has a high awareness of its interdependence with other firms. True myopia in a firm is represented by $\lambda_i = 0$.



Basic Model of Bank Performance (cont.)

Substitution of λ_i in equation 3 gives:

$$p^*Y_i - w_i \frac{dX_i^*}{dY_i} Y_i = -(Y_i)^2 f'(Y)(1 + \lambda_i) \quad (5)$$

Dividing both sides by pY_i and rearranging gives:

$$\frac{p^*Y_i - w_i \frac{dX_i^*}{dY_i} Y_i}{p^*Y_i} = -\frac{Y_i}{Y} \frac{f'(Y)Y}{p^*} (1 + \lambda_i) \quad (6)$$

Now a bank's mark-up can be decomposed into three parts, equivalent to the right-hand side of equation 6:

- 1 (Y_i/Y) is firm i 's market share MS_i , with $0 < MS_i \leq 1$.
- 2 $f'(Y)Y/p$ is the inverse of the price elasticity of demand, $1/\eta$, equal to the market elasticity if and only if all firms are price takers in the output market and $p_i = p, \forall i$.
- 3 $1+\lambda_i$ measures firm i 's expectations about the reactions of its rivals dY/dY_i , with $-1 \leq \lambda_i \leq 1$.



Production theory

Why do we have/need "production theory"?

- to understand and model the "production process"
- to help understand growth processes
- to understand where "waste" or "slack" comes from
- to understand and interpret competition especially for (natural) monopolies
- to be able to learn, adapt, absorb and transform



Basic Model of Bank Performance (cont.)

We can now write equation 6 as:

$$\frac{p^*Y_i - w_i \frac{dX_i^*}{dY_i} Y_i}{p^*Y_i} = (MS_i) \left(-\frac{1}{\eta}\right) (1 + \lambda_i) \quad (7)$$

After multiplying by p^*Y_i we have:

$$\Pi_i^* = p^*Y_i - w_i \frac{dX_i^*}{dY_i} Y_i = (MS_i) \left(-\frac{1}{\eta}\right) (1 + \lambda_i) p^*Y_i \quad (8)$$



Production theory

Elements of a production theory

- A production set: inputs and outputs
- A transformation functions: how to turn inputs into outputs
- A pricing opportunity set: (relative) prices of inputs and/or outputs
- Production dynamics: fixed versus variable costs, and changes in technology



Production sets: defining inputs and outputs

Typical assumptions:

- Production process is monoprotic
- Inputs and outputs are homogenous
- Production function is twice continuously differentiable
- There is no uncertainty regarding the production function and prices
- We know the objectives and constraints of the producer



Transformation functions: efficiency and productivity

"Using" the transformation function:

- Choose your mix of inputs in order to maximize your outputs
- Can you substitute x_1 (e.g. capital) for x_2 (e.g. labor)?
- Important: marginal rate of technical substitution: $\frac{dx_1}{dx_2}$
- Of course the rate at which you want to substitute depends on the MP of each input: $-\frac{MP_{x_1}}{MP_{x_2}}$
- In fact: $\frac{dx_1}{dx_2} = -\frac{MP_{x_1}}{MP_{x_2}}$



Transformation functions: efficiency and productivity

So what is transformation about?

- In the most basic form: $y = f(x_1, x_2)$
- A short run version could read: $y = f(x_1 | x_2 = x_T)$
- Average product (AP) = $y/x_1 = f(x_1 | x_2 = x_T)/x_1$
- Marginal product (MP) = $\frac{\delta y}{\delta x_1} = \frac{\delta}{\delta x_1} f(x_1 | x_2 = x_T)$

Typical production stages:

- 1 increasing AP
- 2 decreasing AP, positive MP
- 3 negative MP



Transformation functions: efficiency and productivity

"Using" the transformation function (cont.):

- Your productivity depends on the amount of inputs you use
- But it also depends on the transformation function
- In particular, it depends on MRTS and on the returns to scale of each input
- For x_1 , the production elasticity is: $\frac{\delta y}{\delta x_1} \frac{x_1}{y}$
- And the total elasticity of production ϵ is the sum of the input elasticities of production
- $\epsilon = 1$ means constant returns to scale
- $\epsilon > 1$ means increasing returns to scale
- $\epsilon < 1$ means decreasing returns to scale



Transformation functions: efficiency and productivity

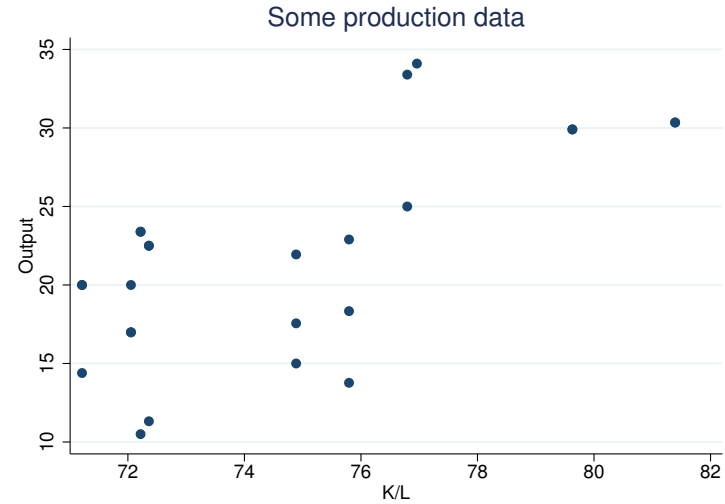
So what if there is slack?

- What if you find that there are two producers, with the same transformation function, and the same inputs, who produce different outputs? How can that be?
 - 1 They pay different prices for their inputs.... We'll come back to that one in a few slides!
 - 2 One is more wasteful, and "spills" some of the inputs: "leakage (heat)", carelessness, different objectives, etc.



Transformation functions: efficiency and productivity

Transformation functions: efficiency and productivity

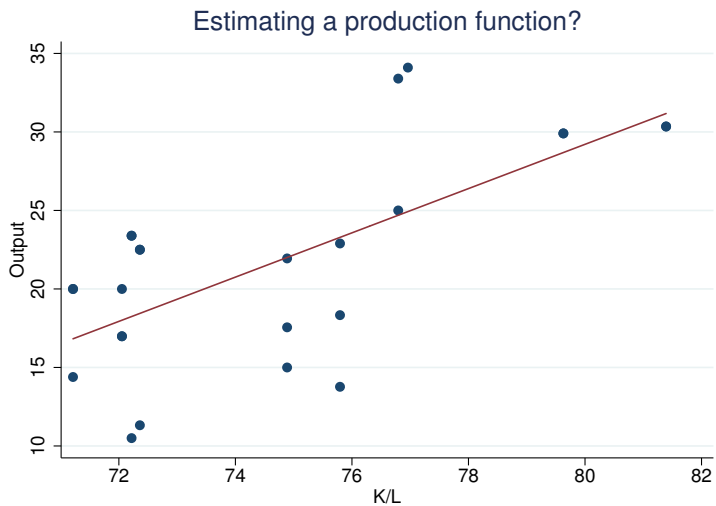


Check out how indeed firms with the same input(s) produce different levels of output

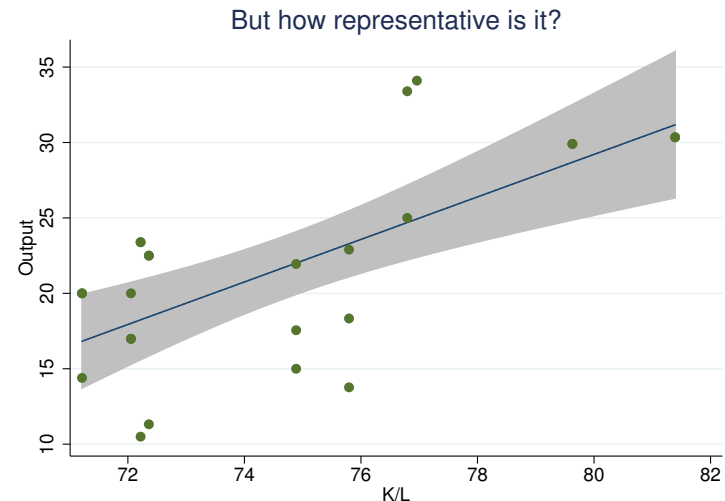


Transformation functions: efficiency and productivity

Transformation functions: efficiency and productivity



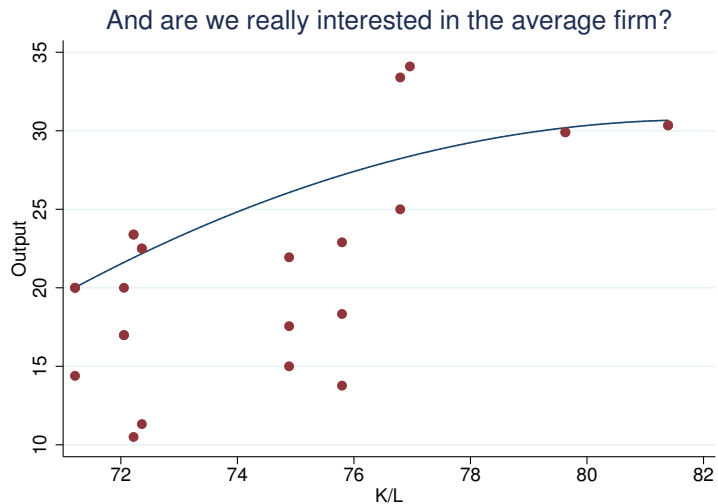
So we estimate something like this: $Output = b_0 + b_1(K/L) + \epsilon$



ϵ is "measurement error"? or slack? or a bit of both?



Transformation functions: efficiency and productivity



So we remove the measurement error from ϵ and keep the slack?



Transformation functions: efficiency and productivity

How do we remove the slack (cont.)?

Some industries, however, may lack the ability to employ existing technologies efficiently and therefore produce less than the frontier output.

If the difference between optimum and actual (observable) output is represented by an exponential factor, $\exp\{-u_{it}\}$, then the actual output, Y_{it} can be expressed as a function of the stochastic frontier output,

$$Y_{it} = Y_{it}^* \exp\{-u_{it}\}$$



Transformation functions: efficiency and productivity

How do we remove the slack?

If all industries produce on the boundary of a common production set that consists of an input vector with two arguments, physical capital (K) and labor (L), output can be described as:

$$Y_{it}^* = f(K_{it}, L_{it}, t; \beta) \exp\{v_{it}\} \tag{9}$$

where:

- Y_{it}^* is the frontier (maximum) level of output of producer i , at time t
- f and parameter vector β characterize the production technology
- t is a time trend variable that captures neutral technical change (more about that later)
- v_{it} is an i.i.d. error term distributed as $N(0, \sigma_v^2)$, which reflects the stochastic character of the frontier.



Transformation functions: efficiency and productivity

How do we remove the slack (cont.)?

Equivalently:

$$Y_{it} = f(K_{it}, L_{it}, t; \beta) \exp\{v_{it}\} \exp\{-u_{it}\} \tag{10}$$

where:

- $u_{it} \geq 0$ is assumed to be i.i.d., with a half-normal distribution truncated at zero $|N(0, \sigma_u^2)|$ and independent from the noise term, v_{it}
- We decompose the residual in equation (10), $\exp\{\epsilon_{it}\} = \exp\{v_{it}\} \exp\{-u_{it}\}$,
- and identify its components, $\exp\{v_{it}\}$ and $\exp\{-u_{it}\}$, by re-parameterizing λ in the maximum likelihood procedure,
- where $\lambda (= \sigma_u / \sigma_v)$, the ratio of the standard deviation of efficiency over the standard deviation of the noise term,
- and $\sigma (= (\sigma_u^2 + \sigma_v^2)^{1/2})$ is the composite standard deviation.



Transformation functions: efficiency and productivity

How do we remove the slack (cont.)?

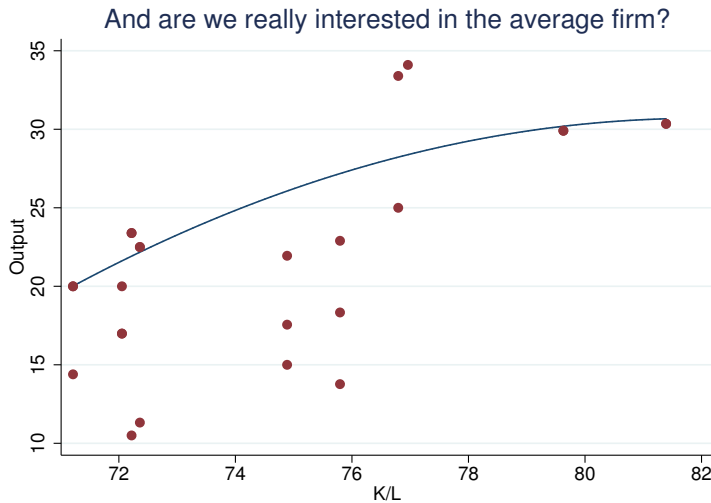
$$Y_{it} = f(K_{it}, L_{it}, t; \beta) \exp\{v_{it}\} \exp\{-u_{it}\} \quad (11)$$

- The frontier is identified by the λ for which the log-likelihood is maximized.
- Technical efficiency, $\exp\{-u_{it}\}$, is measured as the ratio of actual over maximum output, $\exp\{-u_{it}\} = \frac{Y_{it}}{Y_{it}^*}$
- such that $0 \leq \exp\{-u_{it}\} \leq 1$
- and $\exp\{-u_{it}\} = 1$ implies full efficiency.



Transformation functions: efficiency and productivity

Remember this?



Transformation functions: efficiency and productivity

It all makes a lot more sense when we actually estimate it:

To operationalize equation (10), we need to specify the functional form of the production frontier. Let's say the production function looks like this (disregarding the noise for a second):

$$Y_{it} = K_{it}^{\beta_k} L_{it}^{\beta_l} \quad (12)$$

If we estimate this in logs, we get:

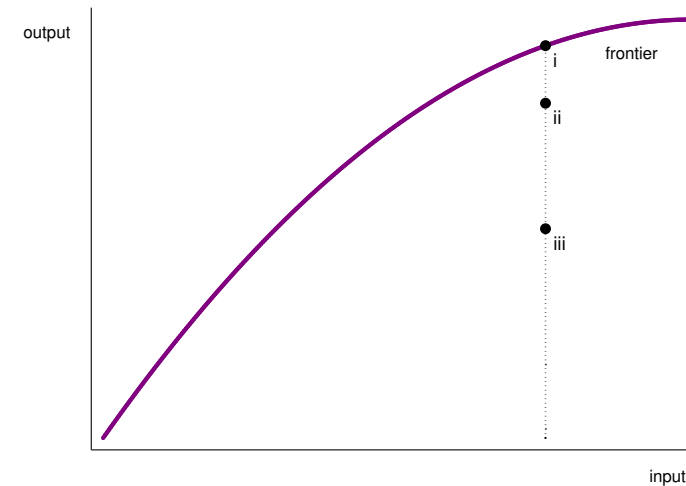
$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + v_{it} - u_{it} \quad (13)$$

where lower case letters denote logarithms.



Transformation functions: efficiency and productivity

Now we can compare (i), (ii), (iii)...



Pricing opportunity sets: technical versus allocative efficiency

- Most common is the measurement of technical efficiency, $\exp\{-u_{it}\}$.
- Inputs ("raw materials") tend to be highly homogenous for a given production technology
- More likely, production technologies are different...
- So we typically ignore "allocative efficiency"
- But what about output prices?



Frontier shifts: (augmented) technical change

Exactly, instead of this:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + v_{it} - u_{it} \quad (16)$$

I should have noted (and estimated):

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_t t + v_{it} - u_{it} \quad (17)$$

Or should I have even done something different?



Frontier shifts: (augmented) technical change

If you watched carefully, you've noticed that I have not been very consistent in my notation ...

Remember I started with his:

$$Y_{it} = f(K_{it}, L_{it}, t; \beta) \exp\{v_{it}\} \exp\{-u_{it}\} \quad (14)$$

And ended up with the following empirical specification:

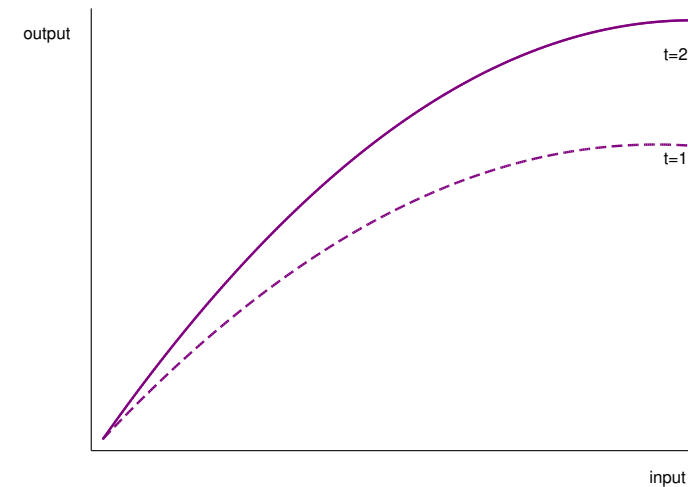
$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + v_{it} - u_{it} \quad (15)$$

where is the mistake?



Frontier shifts: (augmented) technical change

Consider the following situation:



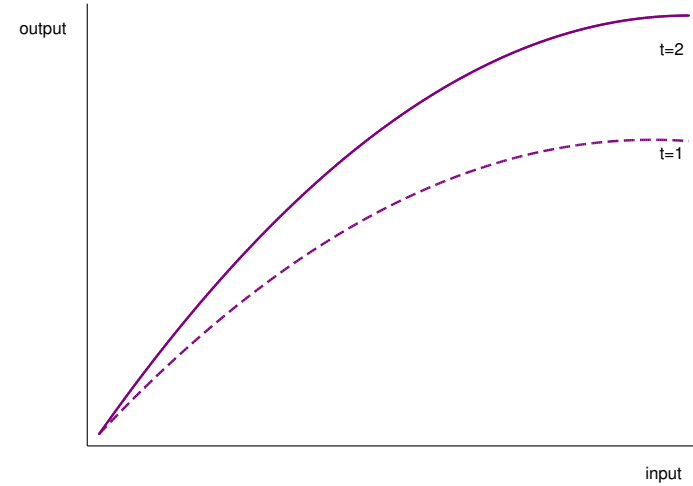
Does β_t capture the shift in the production frontier?



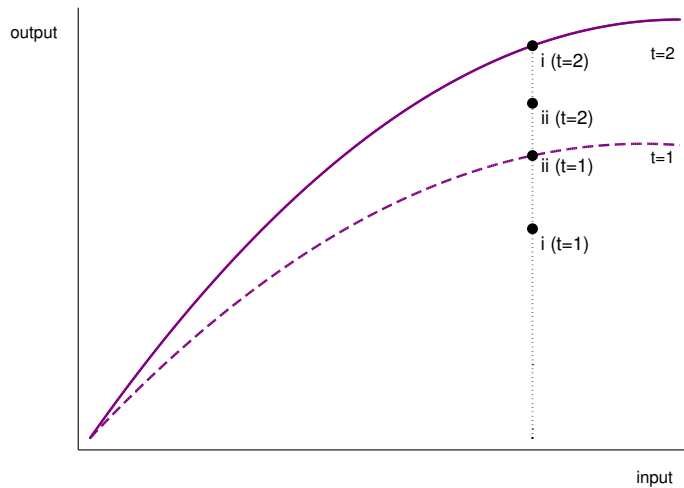
Frontier shifts: (augmented) technical change

Or should I consider:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_t t + \beta_{kt} k_{it} t + \beta_{lt} l_{it} t + v_{it} - u_{it} \quad (18)$$



Frontier shifts: (augmented) technical change



Frontier shifts: (augmented) technical change

The Diamond-McFadden impossibility Theorem

... states that it is impossible to independently identify substitution and technological change in most normal cases.

So what do applied economists have to say in this regard?
The simplest solution:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_t t + v_{it} - u_{it} \quad (19)$$

We already had a look at:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_t t + \beta_{kt} k_{it} t + \beta_{lt} l_{it} t + v_{it} - u_{it} \quad (20)$$



The Diamond-McFadden impossibility Theorem

... states that it is impossible to independently identify substitution and technological change in most normal cases.

So what do applied economists have to say in this regard?

A more flexible approach:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \frac{1}{2} \beta_{kk} k_{it}^2 + \frac{1}{2} \beta_{ll} l_{it}^2 + \beta_{kl} k_{it} l_{it} + \gamma_t D_t + \delta_{kt} k_{it} D_t + \delta_{lt} l_{it} D_t + v_{it} - u_{it} \quad (21)$$

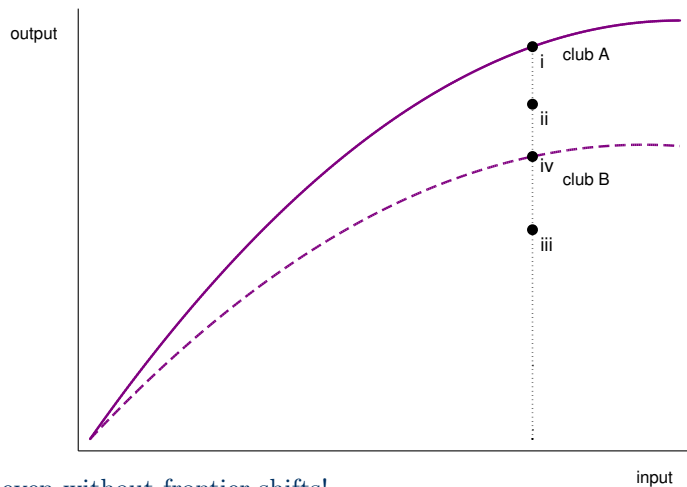
where D are time dummies.

In addition, we could make u_{it} a function of t (or D)



The Diamond-McFadden impossibility Theorem

For $z = 2$, we'd then have:



... even without frontier shifts!



The Diamond-McFadden impossibility Theorem

So let's make matters worse: what if there is more than one "appropriate technology" (transformation function)? For example, z functions:

$$y_{it} = \beta_{o|z} + \beta_{k|z} k_{it} + \beta_{l|z} l_{it} + \frac{1}{2} \beta_{kk|z} k_{it}^2 + \frac{1}{2} \beta_{ll|z} l_{it}^2 + \beta_{kl|z} k_{it} l_{it} + \gamma_{t|z} D_t + \delta_{kt|z} k_{it} D_t + \delta_{lt|z} l_{it} D_t + v_{it|z} - u_{it|z} \quad (22)$$



The Diamond-McFadden impossibility Theorem

So in the end:

- What can an applied economist say about the impossibility theorem?
 - If specifications are nested, you can test which specification is preferred (e.g. Cobb-Douglas versus Translog)
 - Testing for preferred specifications, however, is subject to path dependency
 - However, out-of-sample testing is still very difficult
 - And you often need strong economic priors to arrive at the preferred specification
- And why again is this important?
 - Main reason: contemporaneous efficiency versus long-run technical change
 - Capital investments lower one, but may increase the other
 - What is the optimal trade-off?
 - What is the best (real) option?

▶ skip



Production functions and cost functions

- Production theory is about ... maximizing production
- But what about minimizing costs?
- Or maximizing profits?
- This is where duality comes in handy ...



Production functions and cost functions



Duality: so what good does this do us?

- A producer that maximizes output subject to a cost constraint (input constraint), also minimizes costs subject to an output objective.
- So cost minimization is the dual of output maximization (and therefore shares many of the same properties.
- In perfect competition, a producer that minimizes costs, maximizes profits.
- In perfect competition, profit maximization is the dual of output maximization (and cost minimization)

Production functions and cost functions

Duality:

- Optimal output: $y^* = y^*(p, w_1, w_2)$
- Input demand functions: $x_1^* = x_1^*(p, w_1, w_2)$, $x_2^* = x_2^*(p, w_1, w_2)$
- Profit function: $\pi = TR - TC = py - (w_1x_1 + w_2x_2)$
- Substituting production function into profit function:
 $\pi = TR - TC = p(f(x_1, x_2)) - (w_1x_1 + w_2x_2)$
- Substituting optimal inputs into production function:
 $y = f(x_1, x_2) = f(x_1^*(p, w_1, w_2), x_2^*(p, w_1, w_2))$



Production functions and cost functions



Cost function properties:

- 1 $c^*(y, w) \geq 0$, for $w \geq 0$ and $y > 0$
- 2 $c^*(y, w^a) \geq c^*(y, w^b)$, for $w^a \geq w^b$
- 3 $c^*(y, w)$ is homogenous of degree one in all prices
- 4 $\frac{\partial c^*(y, w)}{\partial w_i}$ is homogenous of degree zero in all prices
- 5 $c^*(y, w)$ is weakly concave in input prices if the production function is strictly quasi-concave

Production functions and cost functions

Profit function properties:

- 1 $\pi^*(p, w) \geq 0$, for $p, w \geq 0$
- 2 $\pi^*(p^a, w) \geq \pi^*(p^b, w)$, for $p^a \geq p^b$
- 3 $\pi^*(p, w^a) \leq \pi^*(p, w^b)$, for $w^a \geq w^b$
- 4 $\pi^*(p, w)$ is homogenous of degree one in all prices
- 5 $\frac{\delta \pi^*(p, w)}{\delta w_i}$ and $\frac{\delta \pi^*(p, w)}{\delta p}$ are homogenous of degree zero in all prices
- 6 $\pi^*(p, w)$ is convex in all prices if the production function is strictly concave



X-Efficiency

- Berger, Hunter and Timme (1993) define X-efficiency as the economic efficiency of any single firm minus scale and scope efficiency effects.
- Economic efficiency is the sum of technical and allocative efficiency.
- Technical efficiency is a measure of a bank's distance from the frontier, minimizing inputs given outputs or vice versa.
- Allocative efficiency measures the extent to which a bank is able to use inputs and outputs in optimal proportions given prices and the production technology.



Efficiency of Banks

- In all models introduced so far, we have assumed that banks choose optimal output prices p and inputs x that maximize profits, given existing market power.
- Therefore, any deviations from the profits that would prevail under perfect competition are entirely attributed to (changes in) the degree of competition in the market.
- In practice, of course, banks may choose suboptimal combinations of output prices and inputs. They may produce output at a suboptimal scale, produce a suboptimal combination of outputs, or select a suboptimal combination of inputs (or input prices) to produce outputs.
- In short, banks may be inefficient.



X-Efficiency (cont.)

- We start by making the basic model stochastic:

$$\Pi_i^* = (p^* Y_i - w_i \frac{dX_i^*}{dY_i} Y_i) * \exp(\varepsilon_i) \tag{23}$$

- We assume that ε_i can be decomposed into a noise component v_i , and an efficiency component u_i , where $\varepsilon_i = v_i - u_i$.
- All inefficient firms will operate below the efficient profit frontier.
- Profit efficiency for bank i is defined as:

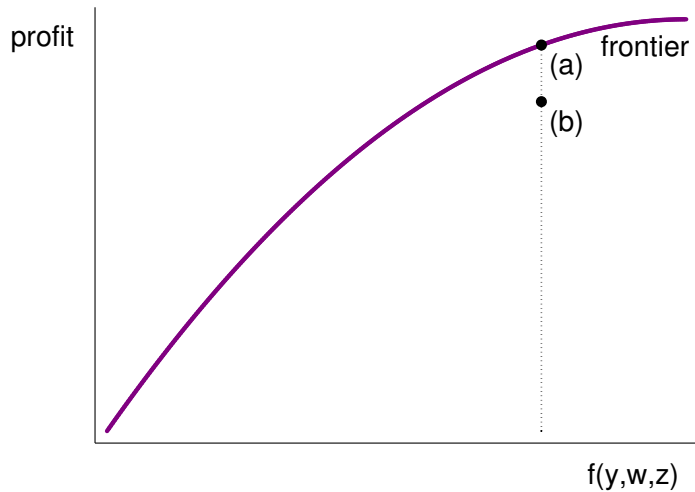
$$PE_i = E[\exp(-u_i) | \varepsilon_i] \tag{24}$$

- This measure takes on a value between 0 and 1, where 1 indicates a fully efficient bank.



X-Efficiency (cont.)

Figure: Profit Frontier



Scale and Scope Economies (cont.)

Scope

- The extent to which that output mix itself is optimal is measured by calculating scope economies.
- Unfortunately, calculating scope economies is not as straightforward as calculating scale economies.
- The derivation itself is straightforward, however, and analogous to Equation 25:

$$\frac{\partial \Pi_i^*}{\partial Y_{i,k}} \frac{\partial \Pi_i^*}{\partial Y_{i,l}}, \text{ for } k \neq l \quad (26)$$

- The main problem with this method lies in the fact that, at least theoretically, we require banks with zero outputs for specific outputs Y_k (cf. Berger and Humphrey (1994)).



Scale and Scope Economies

Scale

- We define output-specific economies of scale as the ceteris paribus increase in profits that results from an increase in output Y_k .
- To this purpose we take Equation 8 and calculate the derivative respect to Y_k :

$$\frac{\partial \Pi_i^*}{\partial Y_{i,k}} \quad (25)$$

- A value larger (smaller) than one indicates increasing (decreasing) returns to scale, and unity indicates constant returns to scale.
- Overall economies of scale are simply the sum of output-specific economies of scale.



Scale and Scope Economies (cont.)

Scope (cont.)

- The main problem with this method lies in the fact that, at least theoretically, we require banks with zero outputs for specific outputs Y_k (cf. Berger and Humphrey (1994)).
- The models we have discussed so far are usually estimated using logarithmic (semi-)flexible forms and thereby cannot handle these zero outputs.
- In addition, Berger, Hanweck, and Humphrey (1987) observed that for translog functions complementarities cannot exist at all levels of output.
- Finally, in many cases there is an extrapolation problem as well: Given a sample containing both universal banks and other banks, only the former typically offer the full range of financial services. Consequently, the economies of scope derived from the cost (or profit) function tend to overestimate the true economies of scope among most sample banks.



Scale and Scope Economies (cont.)

Scope (cont.)

- An alternative method is suggested in Bos and Kolari (2005):
 - 1 Specify a model with three outputs, Y_1 - Y_3 , which sum to Y .
 - 2 Define $Y_1/Y = a$, $Y_2/Y = b$ and $Y_3/Y = c$.
 - 3 Calculate $d = a^2 + b^2 + c^2$. This measure is bounded by $1/3$ (not specialized) and 1 (specialized).
 - 4 Define 'high' [H] as referring to the upper 25th percentile, and 'low' [L] for the remainder of the observations.²
 - 5 Now, the ratio $(\Pi_L^* - \Pi_H^*) / \Pi_H^*$ can be calculated for Y_1 - Y_3 , and Y .
 - 6 Profits Π_i^* are divided by total revenues to adjust for the possibility that banks in high and low bank groups may be different in size.
 - 7 If scope economies exist, the ratio is greater than 0.
 - 8 Bos and Kolari (2005) report a t-value for an independent samples test for $\Pi_L^* - \Pi_H^*$.

²Note that by varying the cut-off point above and below the 25th percentile, it is possible to check for extrapolation problems.



Synthesis

Model Hypotheses	Key assumptions	Key variable(s)
PE $(1 - \frac{\Pi_i^*(Y_i, w_i) * \exp(v_i)}{\Pi_i^*(Y_i, w_i)}) > 0$	η constant, λ_i and $MS_i = f(p)$	$\varepsilon_i = v_i - v_i$
Scale $\frac{\partial \Pi_i^*(Y_i, w_i)}{\partial Y_{i,k}} > 0$	η , λ and MS_i absent	Y_i
Scope $\frac{\partial \Pi_i^*(Y_i, w_i)}{\partial Y_{i,k} \partial Y_{i,l}} > 0$ for $k \neq l$	η , λ and MS_i absent	$Y_{i,k}, Y_{i,l}$

