

Advanced Industrial Organization I

Lecture 4: Technology and Cost

Måns Söderbom*

3 February 2009

*Department of Economics, University of Gothenburg. Office: E526. E-mail: mans.soderbom@economics.gu.se

1. Introduction

References for this lecture:

- These notes.
- Chapter 4 in Pepall et al. (2008)
- Akerberg, D., L. Benkard, S. Berry, and A. Pakes (2006), “Production Functions,” Section 2 of “Econometric Tools for Analyzing Market Outcomes” forthcoming in Handbook of Econometrics, Volume 6.
- Griliches Z. and J. Mairesse (1998). ‘Production functions: The Search for Identification,’ in Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium, 169-203. Cambridge University Press.
- Olley, S. and Pakes, A. (1996) ‘The dynamics of productivity in the telecommunications equipment industry’, *Econometrica* 64, 1263-1297.

The latter 3 references can be obtained from the course web-page.

- In the first part of this lecture I will discuss basic the relationship between technology, cost and market structure. In the second part of the lecture I will discuss estimation of production functions.

2. Technology, Cost & Implications for Market Structure

Reference: Chapter 4 in Pepall et al. (2008)

- We saw in Lecture 3 that the shape of the **demand** function is important for the decisions firms are making, and for the market structure. For example, for a perfectly competitive market, shifts in the demand curve may affect profits in the short term and the number of firms in the market in the long run.
- **Production costs** are another important factor explaining firm behaviour and industry structure. We saw in Lecture 3 how the optimal quantity supplied by a single firm to the market likely depends on the marginal cost.
- An important factor determining production costs is the firm's **technology**.
- Firm's technology = a production relationship that describes how a given quantity of **inputs** is transformed into the firm's **output**. In (neoclassical) models, the firm's technology is typically represented by a **production function**, e.g.

$$q = f(x_1, x_2, \dots, x_k),$$

where q is output, x_1, x_2, \dots, x_k are inputs (e.g. labour, capital, raw materials, electricity,...), and $f(\cdot)$ is the production function.

- As you know, the production function is often assumed to be Cobb-Douglas, and the most basic specification is one where there are two inputs, namely physical capital (K) and labour (L), and one additional term A which we shall refer to as total factor productivity:

$$q = AK^\alpha L^\beta,$$

in logs:

$$\ln q = \ln A + \alpha \ln K + \beta \ln L.$$

- Not a very good description of what happens inside the firm - but as a cornerstone of a model of how some of the firm's decisions depend on demand and costs, it will be OK in most cases.
- In IO, we want to understand the firms' input and output decisions for several reasons - e.g. because they will impact on the structure of the market, hence possibly on the level of competition and thereby consumer welfare. To understand these decisions, we need to have a good understanding of the firm's technology.
- Consider a simple model in which the firm chooses output q and inputs x_1, x_2, \dots, x_k in order to maximize profits. To find the firm's optimal level of output and inputs, we proceed in stages:

1. Consider a specific level of output \bar{q} . Suppose that the firm were to choose precisely this level of output - what would the associated cost be? Total cost, by definition, is equal to

$$\text{cost} = \sum_{i=1}^k w_i x_i$$

where w_i is the unit cost of the i th input. Hence, in order to establish total cost, we need to know the selection of inputs (capital, labour,...), and the factor prices. Since we assume firms seek to maximize profits, it follows that, conditional on output, the firm will choose inputs x_1, x_2, \dots, x_k so as to **minimize** total costs. Conditional on output, we can thus formulate the profit maximization problem as

$$C = \min_{x_1, x_2, \dots, x_k} \sum_{i=1}^k w_i x_i,$$

subject to

$$\bar{q} = f(x_1, x_2, \dots, x_k).$$

2. Now solve the cost minimization problem above, and obtain the optimal levels of inputs, at this particular level of output. These inputs can be written as functions of

1. the level of output
2. the factor prices (wages, cost of capital etc.):

$$x_1 = g_1(w_1, \dots, w_k, \bar{q})$$

$$x_2 = g_2(w_1, \dots, w_k, \bar{q})$$

(...)

$$x_k = g_k(w_1, \dots, w_k, \bar{q})$$

3. Now plug in these expressions into the definition of total costs, introduced above:

$$C(w_1, \dots, w_k, \bar{q}) = \sum_{i=1}^k w_i g_i(w_1, \dots, w_k, \bar{q}).$$

- We now have a **cost function**, telling us how total costs vary with factor prices, and with output - where it is assumed that inputs are always chosen optimally. All that remains now is to choose the optimal level of output. And that, as you know, is the level that maximizes profits, i.e. total revenue minus total costs. As we saw in Lecture 3, in the simple competitive model, this implies output price = marginal cost.

- [Discuss in class] **Example:** Cobb-Douglas (see Pepall et al. Section 4.5). Derive the cost function (4.12), i.e.

$$C = \left[\left(\frac{\alpha}{\beta} \right)^{\frac{\beta}{\alpha+\beta}} + \left(\frac{\beta}{\alpha} \right)^{\frac{\alpha}{\alpha+\beta}} \right] r^{\frac{\alpha}{\alpha+\beta}} w^{\frac{\beta}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}}.$$

Discuss how $\alpha + \beta$ determines the shape of the cost function. Formulation suitable for empirical

analysis:

$$\ln C = \ln \left[\left(\frac{\alpha}{\beta} \right)^{\frac{\beta}{\alpha+\beta}} + \left(\frac{\beta}{\alpha} \right)^{\frac{\alpha}{\alpha+\beta}} \right] + \frac{\alpha}{\alpha+\beta} \ln r + \frac{\beta}{\alpha+\beta} \ln w + \frac{1}{\alpha+\beta} \ln q,$$

indicating that it will be possible to estimate α and β by running a regression in which log total costs is the dependent variable, and where log output and log input prices are the explanatory variables:

$$\ln C = \text{Constant} + \delta_1 \ln r + \delta_2 \ln w + \delta_3 \ln q.$$

In applied work, a more flexible functional form is sometimes used - known as the translog cost function - which includes squares and interaction terms of $\ln r$, $\ln w$ and $\ln q$. We don't have to worry about that here - read the last part of Section 4.5 if you are interested.

- However, partly because product specific cost data are not available for many markets, the direct estimation of cost functions has not been an active area of research lately (see discussion in Akerberg, Benkard, Berry and Pakes, 2006, Section 2). Instead, researchers now often attempt to estimate the production function directly, which requires data on inputs and outputs but not on costs. In the second half of this lecture I discuss recent developments in this area of research. In any case, whether you estimate the cost function or the production function, you are essentially looking for the same thing - e.g. evidence on the returns to scale.
- Now let's return to costs. It is useful to distinguish between **fixed costs**, **variable costs**, and **sunk costs**, as follows:
 1. **Fixed cost**, denoted F . Does not depend on output. Describes a given amount of expenditure that the firm must incur in **each period**, regardless of the level of output - e.g. licence fee. Up until now we have ignored fixed costs ($F = 0$).

2. **Variable cost**, denoted $C(q)$. Depends on output. Define:

$$\begin{aligned}\text{Average cost} &= AC(q) = \frac{C(q)}{q} \\ \text{Marginal cost} &= MC(q) = \frac{\partial C(q)}{\partial q} = C'(q)\end{aligned}$$

3. **Sunk costs**. You only pay sunk costs once, and, once paid, it cannot be fully recovered. For example entry costs.

2.1. Cost and output decisions

- We have previously discussed how, in a model with an upward sloping supply function and a downward sloping or flat demand curve, profit maximization in any one period implies that output will be chosen so as to satisfy the following first-order condition:

$$\text{marginal revenue} = \text{marginal cost},$$

which we typically write as

$$MR(q) = MC(q)$$

- This applies in a world in which there are no fixed costs or sunk costs. However, in the presence of fixed and/or sunk costs, the firm has to think through whether it makes sense to produce **anything at all**.
- For example, if fixed costs are so high so as to lead to a **loss** even at the level of output that satisfies the $MR = MC$ condition, it is optimal for the firm to exit from the market (in which case it pays no fixed cost, obviously). So with fixed costs added, we should modify the rule determining output as follows:

- Provided that $pq - C(q) \geq 0$ (or, in other words $p \geq AC(q)$), optimal output satisfies

$$MR(q) = MC(q).$$

- But if $pq - C(q) < 0$, for any $q > 0$, then $q = 0$ is optimal, since the firm prefers zero profit to a loss.
- If there is a sunk entry cost, the potential entrant carries out a similar calculation: if the (expected) present discounted value of all future profits exceeds the sunk entry cost, then the firm will enter the market; otherwise it won't. Once it has entered, the sunk cost doesn't affect the firm's decisions.

2.2. Costs and market structure

- Now consider the relationship between marginal costs, $MC(q)$, and average costs $AC(q)$. The following statement will **always** be true:

Average cost falls whenever marginal cost is less than average cost, and rises whenever marginal cost exceeds average costs.

Intuition: If MC is **low**, then producing one more unit of output raises total cost by **less** than total output, hence AC must fall; if MC is **high**, then producing one more unit of output raises total cost by **more** than total output, hence AC must increase.

- This property of the cost function implies that we have

$$MC(q) = AC(q)$$

precisely at the point where $AC(q)$ is at its **minimum**.

- Why? Because as long as $MC(q) < AC(q)$, an increase in q will reduce $AC(q)$; but as soon as $MC(q) > AC(q)$, an increase in q will increase $AC(q)$. Hence at $MC(q) = AC(q)$, the $AC(q)$, which has been falling at each previous level of q will now turn back up again.

[Example: Cobb-Douglas cost function and fixed cost]

- Now, the firm has to break even for it to be worth supplying any output to the market. In view of this, both $AC(q)$ and $MC(q)$ play a role in determining the firm's decisions and thus the market structure.
- Economies of scale: $AC(q)$ falls as output increases
- Diseconomies of scale: $AC(q)$ rises as output increases
- The presence of fixed costs gives rise to economies of scale. Other processes too - see discussion in book, pp. 64-65.

- The fact that scale economies are measured by a falling average cost gives us a precise way to measure their presence.
- *Index of the extent of scale economies:*

$$S = \frac{AC(q)}{MC(q)}.$$

Notice that

$$\begin{aligned} S &= \left[\frac{MC(q)}{AC(q)} \right]^{-1} \\ S &= \left[\frac{dC(q)/dq}{C/q} \right]^{-1} \\ S &= \left[\frac{dC(q)}{dq} \frac{C}{q} \right]^{-1} \\ S &= [\eta_c]^{-1}, \end{aligned}$$

where η_c is the output elasticity of cost - i.e. it measures the percentage increase in cost resulting from a one percent increase in output. The index S is the **inverse** of the output elasticity of cost.

- If you get little additional output as a result of a large increase in cost, so that $\eta_c < 1$, then $S > 1$, in which case there are diseconomies of scale.
- If you get a lot of additional output as a result of a small increase in cost, so that $\eta_c > 1$, then $S < 1$, in which case there are economies of scale.
- Define **minimum efficient scale** as the lowest level of output at which economies of scale are exhausted, i.e. at which $S = 1$.
- Relationship between scale economies and industry structure. For example, if demand is low and fixed costs are high, the market may be a natural monopoly: only one firm can break even (or do better) in such a market; were there two firms in the market, they would both make a loss (since they both have to incur the fixed cost). Scale economies tend to imply highly concentrated markets:

it makes sense for one or a few firms to produce all the output for the market.

2.3. Sunk cost and market structure

- Sunk costs also affect the market structure. Faced with sunk entry costs, firms will only enter a market if they believe they can at least break even.
- This implies that the present discounted value of future expected profits must exceed the entry cost for the firm to enter the market.
- So sunk costs affect the long-run equilibrium: firms will stop entering the industry when the sunk entry cost exceeds the present discounted value of profits. Everything else equal, the more firms that enter an industry, the more competitive its pricing is likely to be. Hence high sunk costs may result in high prices, potentially harming consumer welfare.
- **Example** (see Section 4.2 in the book).
- Suppose the elasticity of demand in a particular market is equal to 1:

$$\eta = 1.$$

- This implies that consumer expenditure, denoted E , for the product is **constant**:

$$E = P \times Q,$$

since a 1% increase in price is always associated with a 1% fall in output.

- Total output Q is equal to the output of each firm times the number of firms:

$$Q = N \times q.$$

- Hence

$$q = \frac{E}{N \times P}.$$

- Suppose the **Lerner index**, a measure of the extent of monopoly power in the industry (see Lecture 3), depends on the number of firms N as follows:

$$LI = \frac{P - c}{P} = \frac{A}{N^\alpha},$$

where $A > 0$ and $\alpha > 0$ are constants (A does **not** mean total factor productivity in this particular example!).

- Assume firms operate in one period only, so that break even requires that

$$(P - c)q = F,$$

where F is the sunk cost.

- We can now solve for the equilibrium number of firms in the market, denoted N^e . First, note that the break even condition (recall that the long run equilibrium is defined as the point where firms break even) implies

$$(P - c) \frac{E}{N^e \times P} = F.$$

Hence

$$\begin{aligned} \frac{(P - c)}{P} \frac{E}{N^e} &= F \\ \frac{A}{(N^e)^\alpha} \frac{E}{N^e} &= F \\ \left(\frac{AE}{F} \right) &= (N^e)^{1+\alpha}, \end{aligned}$$

thus

$$N^e = \left(\frac{AE}{F} \right)^{\frac{1}{1+\alpha}}.$$

Since $\alpha > 0$, this implies that a high sunk cost, relative to consumer expenditure, will be associated

with a small number of firms in the market, i.e. a highly concentrated market.

You may skip Section 4.3 "Costs and multiproduct firms".

2.4. Noncost determinants of industry structure

We have discussed the implication of cost relationships for market structure. Of course, there are other factors that may play a role in this context. Pepall et al. (Section 4.4) discuss the three such factors.

Market size and competitive industry Suppose the fixed cost of operating a business is quite large. This implies that the minimum efficient scale is also quite large; that is, firms in such an industry will be quite large. Does this then imply that the market is highly concentrated? Not if the market itself is large.

This begs the question: "How big must a market have to be in order to avoid domination by a few big firms?". The rather general answer to this question is: "the more extensive are scale economies, the larger the market has to be". That is, the relationship between market structure and market size will vary according to the market being examined.

- If scale economies are exhausted at some point ($S = 1$), and if sunk entry costs are **constant**, then market concentration should decline as market size grows.
- Intuition: In large markets, there is simply room for more firms - and the reason existing firms don't expand further is that they hit increasing average costs (and so it may not be in their interest to grow). This appears to be the case in Swedish food retailing: recall from assignment 1 that large local markets are less concentrated than small local markets
- [Show graph based on Swedish food retailer data]
- However if sunk or fixed costs are endogenous, the market concentration may not necessarily be low even if the size of the market is large. Pepall et al. (pp.75-76), citing Sutton, provide an example in which the sunk cost F depends on market size:

$$F = K + \beta \times AE.$$

Why might this be? Maybe because of advertising or R&D expenditures - these arguably need to

be bigger in larger markets. Recall our equation for the equilibrium number of firms:

$$N^e = \left(\frac{AE}{F} \right)^{\frac{1}{1+\alpha}}.$$

With endogenous sunk costs, this becomes

$$\begin{aligned} N^e &= \left(\frac{AE}{K + \beta \times AE} \right)^{\frac{1}{1+\alpha}} \\ N^e &= \left(\frac{1}{\frac{K}{AE} + \beta} \right)^{\frac{1}{1+\alpha}}. \end{aligned}$$

Even if the market size, represented by E , grows to infinity, the number of firms may be quite small.

If $\beta = 0.0625$ and $\alpha = 1$, for example, then the equilibrium number of firms in this market cannot be larger than 4.

Network externalities and market structure Network externalities = a consumer's willingness to pay for a product increases as the number of other consumers of the same product rises. Example: telephone system, computer software (e.g. MS Word). Markets for products associated with network externalities are likely to be highly concentrated.

Government policy Regulation of markets may affect the structure of the market - e.g. cap on number of taxis in town. Such policies often result in higher concentration than what would be the case in a free market.

3. Estimation of the production function

References:

Section 2 ("Production Functions") in Akerberg, Benkard, Berry and Pakes (2006).

Griliches Z. and J. Mairesse (1998).

In this section we discuss direct estimation of the production function. We focus on the simple 2-factor Cobb-Douglas production function, which we now write as

$$Y_j = A_j K_j^{\beta_k} L_j^{\beta_l},$$

where Y is output (or value-added), A is total factor productivity, K is capital, L is labour, and β_k, β_l are parameters.

3.1. Why are we interested?

- In so far as there is one thing on which economists appear to be able to agree it is the desirability of higher **productivity**. The production function is an important tool that can be used to analyze various aspects of productivity. Here are some research questions/issues that can be addressed using a production function approach:
- **Scale and productivity.** In most datasets, labour productivity (usually defined as output or value-added per worker) is much higher large than small firms. We will see this in the dataset that underlies assignment 2. Is this because large firms have more capital per worker, or because there are increasing returns to scale? If we believe the production function above is correctly specified, we can answer this question by estimating β_k and β_l .

[Show graph of $\log Y/L$ against $\log K/L$, based on the data for assignment 2]

- Suppose we convince ourselves there are increasing returns to scale, i.e. $\beta_k + \beta_l > 1$. One implication would be that if a fixed set of inputs (at the national level) gets allocated to a small number of large firms this results in more aggregate output than if allocated to a large number of small firms. This may be important for policy. It may also help us understand the market structure (cf. the discussion above - increasing returns tend to yield high concentration).
- In contrast, if we convince ourselves returns to scale of constant, $\alpha + \beta = 1$, a reallocation of resources between firms of differing size may not impact on aggregate output (e.g. two small firms will produce as much output as one large firm using the same amount of inputs as the two small ones between them).
- Other potentially interesting questions:
 - Rates of technological change: add time effects to the specification.
 - Rates of return on, for example, R&D or exporting ('learning-by-exporting'): add such variables to the specification (we will look at R&D later in the course).

The contribution of various forms of inputs to output - e.g..skilled & unskilled labour, distinguish different types of labour and estimate the associated parameters.

3.2. Can we estimate by OLS? Probably not.

- We write our production function

$$Y_j = A_j K_j^{\beta_k} L_j^{\beta_l},$$

in logarithmic form,

$$y_j = \beta_0 + \beta_k k_j + \beta_l l_j + \epsilon_j,$$

where

$$\ln A_j = \beta_0 + \epsilon_j$$

is log TFP. β_0 is a constant, interpretable as the mean of log TFP, while ϵ_j measures the deviation in productivity from the mean, for firm j .

- Important: TFP is typically assumed **unobserved** (at least partially) by the researcher, but **observed** by the manager of the firm. In other words, the manager knows more about the productivity of the firm than the person running the regressions does, and the manager will make its decisions partly based on this piece of information that you don't have.
- Suppose we have micro data on output, capital and labour. How can the parameters of the production function be estimated?
- As you know, for OLS to consistently estimate the β -parameters, the error term must have zero mean and be uncorrelated with the explanatory variables:

$$E(\epsilon_j) = 0,$$

$$\text{Cov}(k_j, \epsilon_j) = 0, \tag{3.1}$$

$$\text{Cov}(l_j, \epsilon_j) = 0 \tag{3.2}$$

The zero mean assumption is innocuous, as the intercept β_0 would pick up a non-zero mean in ϵ_j .

- The crucial assumption is zero covariance. Is this likely to hold in the present context?
- No - because it seems quite possible that the firm's capital and labour decisions are influenced by factors that are observed to the firm's manager but unobserved to the econometrician, i.e. by ϵ_j . This would set up a correlation between the regressors and the residuals, rendering the OLS estimates biased and inconsistent.

3.3. Illustration

Assumptions:

- Firms operate in perfectly competitive input and output markets (so that input and output prices are not affected by the actions of firm j);
- Capital is a fixed input (decided upon one period in advance, say) rented at rate r ;
- Firms observe ϵ_j before hiring labour (at rate W), and labour is a 'flexible input' that can be altered without dynamic implications.

The firm's profit is given by

$$\begin{aligned}\pi_j &= pY_j - WL_j - rK_j \\ \pi_j &= p \left(A_j K_j^{\beta_k} L_j^{\beta_l} \right) - wL_j - rK_j,\end{aligned}$$

where p is the output price. Assuming the firm maximizes profits, it will choose labour such the following first-order condition is fulfilled:

$$\beta_l p A_j K_j^{\beta_k} L_j^{\beta_l - 1} = W,$$

which implies

$$L_j = \left(\frac{\beta_l p A_j}{W} \right)^{\frac{1}{1-\beta_l}} K_j^{\frac{\beta_k}{1-\beta_l}},$$

or, in logs,

$$l_j = \frac{1}{1-\beta_l} [\ln \beta_l + \ln p - \ln W + \ln \beta_0 + \epsilon_j + \beta_k k_j].$$

- Clearly in this case l_j - optimal labour - **depends on** unobserved TFP (which is the interpretation assigned to the residual ϵ_j) and so estimating the production function

$$y_i = \beta_0 + \beta_k k_j + \beta_l l_j + \epsilon_j.$$

by means of OLS will give biased and inconsistent results.

- Since the first-order condition for labour implies a positive correlation between l_j and ϵ_j , we would expect the OLS estimate of β_l to be upward biased.
- There are other reasons OLS estimates may not be reliable too. **Attrition** is one such mechanism (Olley and Pakes, 1996). To illustrate, suppose the probability that the firm will **exit from the market** if the value of the firm falls below some threshold Ω :

$$\Pr(\text{exit}_{j,t+1} = 1 | \epsilon_j, k_j) = \Pr(V_j(\epsilon_j, k_j) < \Omega).$$

Suppose further that the value of the firm is an increasing function on unobserved productivity and the level of capital stock installed. In this case, the typical firm that would exit would be one with

- a low level of productivity; and
- a low level of capital

(this type of firm would have a low value).

- Think about what this means for the correlation between unobserved productivity and observed capital in **your sample** of survivors.
- Firms with a lot of capital are likely to survive even if they have low productivity, because they have high values.’
- However firms with little capital will only survive if they have high levels of productivity.
- Hence, in the *sample of survivors* there will be a **negative** correlation between k_j and unobserved productivity ϵ_j .
- Thus, if we estimate the production function

$$y_j = \beta_0 + \beta_k k_j + \beta_l l_j + \epsilon_j,$$

this mechanism would tend to yield a downward bias in the coefficient on k_j .

4. Traditional solutions to the endogeneity problem

The two traditional solutions to endogeneity problems are **instrumental variables** and **fixed effects**.

We are now going to write the production function as

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + (\omega_{jt} + \eta_{jt}),$$

i.e. we have added time subscripts assuming we have panel data; and we have decomposed the residual ϵ into two components, $\omega_{jt} + \eta_{jt}$

- ω_{jt} represents the part of TFP observable to the firm but not to the econometrician - hence this is the source of endogeneity problems. You can think of ω_{jt} as a measure of the managerial quality of the firm. From now on, we will refer to ω_{jt} as **unobserved productivity**.
- η_{jt} on the other hand is assumed not to impact on the firm's input decisions. You can think of η_{jt} as representing measurement errors in output, for example (other interpretations are possible too; see Section 2.2 in Akerberg et al.). What's important is that η_{jt} is not a source of endogeneity bias.

4.1. Instrumental Variables

Our problem: We want to estimate

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + \omega_{jt} + \eta_{jt},$$

but we cannot use OLS, since

$$\text{Cov}(l_{jt}, \omega_{jt}) \neq 0.$$

(It is likely, of course, that capital is endogenous too, but we abstract from that possibility for the moment.)

Suppose an instrument z_{jt} is available, that fulfills the following conditions:

1. The instrument is **valid** (or **exogenous**):

$$\text{cov}(z_{jt}, \omega_{jt}) = 0.$$

This is an **exclusion restriction** - z_{jt} is excluded from the structural equation (the production function).

2. The instrument is **informative** (or **relevant**). This means that the instrument z_{jt} must be correlated with the endogenous regressor (labour in the current example), conditional on all exogenous variables in the model (i.e. capital, if this is thought exogenous). That is, if we assume there is a linear relationship between l_{jt} and z_{jt} and k_{jt} ,

$$l_{jt} = \delta_0 + \delta_1 k_{jt} + \theta_1 z_{jt} + r_{jt}, \tag{4.1}$$

where r_{jt} is mean zero and uncorrelated with the variables on the right-hand side, we require $\theta_1 \neq 0$.

Many economists take the view that, for instrumental variable estimation to be convincing, the instruments used must be motivated by theory. Recall the first-order condition for labour derived above -

with my slightly modified notation we get

$$l_{jt} = \frac{1}{1 - \beta_l} [\ln \beta_l + \ln p - \ln W + \beta_k k_{jt} + \omega_{jt}].$$

- This suggests the wage rate W might be a useful instrument:
 - Our theory says it is (negatively) correlated with labour.
 - The wage rate also must be uncorrelated with ω_{jt} . This may not be an entirely innocuous assumption to make. While the wage rate does not directly enter the production function, wages might be correlated with unobserved productivity for other reasons - e.g. if more productive firms have stronger market power in input markets - in which case the wage will not be a valid instrument.

- It also follows from the first-order condition above that the output price is a potential instrument - however, that has been used less often in the literature. Why might we be concerned about using the output price as an instrument?

- A similar way of reasoning can be applied for capital, if that is thought endogenous (i.e. use the cost of capital as an instrument).

Five reasons why the IV approach based on prices as instruments has not been very successful

1. **Market power.** Wages and capital prices (and output prices) could well be correlated with unobserved productivity if input (output) markets are not perfectly competitive: e.g. high unobserved productivity gives the firm market power and so enables it to influence the price.
2. **Wages and unobserved worker quality.** When labour costs are reported in firm-level datasets, they typically come in the form of average wage per worker, and you may well be concerned that the average wage in the firm is correlated with unobserved quality of the workforce. Since the unobserved quality of the workforce likely impacts on unobserved productivity, this would imply the average wage is an invalid instrument.
3. **Law of one price.** If, as is typically the case, one wants to include time dummies in the production function, there must be variation in input prices **across** firms at a given point in time for these to be useful instruments. If input markets are essentially national in scope, this seems unlikely. (If average wages indeed vary across firms in most datasets, you suspect this is at least partly picking up unobserved worker quality).
4. **Endogenous unobserved productivity.** Suppose unobserved productivity ω_{jt} actually **depends** on input choices - e.g. investment in modern technology raises productivity. In that case it will be hard to argue that input prices are valid instruments, since these surely will impact on investment.
5. **Attrition.** A different kind of endogeneity problem sometimes discussed in the literature is posed by endogenous attrition (see above), i.e. that the firm's exit decision depends on unobserved productivity as well as input prices (after all, these jointly determine the profitability of the firm). In such a case input prices cannot be used as instruments.

The common theme across these reasons is that prices are likely to be **invalid instruments**. Recall,

we need

$$\text{cov}(z_{jt}, \omega_{jt}) = 0$$

to hold for the IV estimator to work, if the production function is:

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + (\omega_{jt} + \eta_{jt}).$$

If this condition does not hold - for any of the five reasons just discussed - so that

$$\text{cov}(z_{jt}, \omega_{jt}) \neq 0$$

we say that the instrument z_{jt} is invalid, implying that the IV approach won't work.

4.2. Fixed Effects

A second traditional solution to the endogeneity problem is **fixed effects** estimation, which requires **panel data**. One key assumption underlying this approach is that unobserved productivity is **constant** over time,

$$\omega_{jt} = \omega_j$$

but varies across firms. We would now write the production function as

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + (\omega_j + \eta_{jt}),$$

and use the within estimator ('fixed effects' estimator) to estimate the parameters.

The source of endogeneity bias is now controlled for, thus effectively solving the endogeneity problem - provided, of course, unobserved productivity really **is** constant over time.

The fixed effects approach has not been entirely successful in practice. Two main reasons:

1. **Time invariant unobserved productivity.** The assumption that unobserved productivity is fixed over time is thought unrealistic, especially in longer panels.
2. **Poor performance in practice.** Fixed effects estimates of the capital coefficient are often implausibly low, and estimated returns to scale is often (severely) decreasing ($\beta_k + \beta_l \ll 1$).

4.3. The Olley and Pakes (1996) approach

Note: This is optional material. Only read this if you have time.

The Olley & Pakes (1996; henceforth OP) use a different approach for solving the endogeneity problems discussed above. Similar to the IV approach, OP derive their solution from the **input demand equations**, however OP do not require factor prices to be observed. In what follows I will discuss a simplified version of the OP model.

- The production function:

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_l l_{jt} + (\omega_{jt} + \eta_{jt}).$$

(the original OP model also allows for an effect of firm age, but I ignore that here; I also ignore the possibility acknowledged by OP that endogenous exit - attrition - may cause bias).

4.3.1. Summary of key assumptions

- **Labour** is a flexible input chosen in period t , after observing productivity ω_{jt} .
- **Capital** is a "quasi-fixed" input chosen in period $t - 1$ and evolves according to the equation

$$K_{jt} = (1 - \delta) K_{j,t-1} + I_{j,t-1},$$

where $I_{j,t-1}$ denotes investment.

- Unobserved **productivity** ω_{it} exhibits first-order serial correlation, so that firms with a relatively high productivity today are likely to have a relatively high productivity tomorrow.¹ For example,

¹Strictly speaking, it is assumed that unobserved productivity follows a first order Markov process,

$$p(\omega_{j,t+1} | \{\omega_{j\tau}\}_{\tau=0}^t, I_{jt}) = p(\omega_{j,t+1} | \omega_{jt}),$$

where I_{jt} is the firm's information set in period t . This means that, given the present information, future states are independent of the past states - lags of the productivity variable do not provide additional information as to what might happen to productivity in the future. If you find this terminology confusing, just ignore this footnote.

unobserved productivity may follow a first order autoregressive process:

$$\omega_{jt} = \rho\omega_{j,t-1} + \xi_{jt}.$$

- The **profit** in period t is defined as

$$\Pi_t = pK_{jt}^{\beta_k} L_{jt}^{\beta_l} \exp(\beta_0 + \omega_{jt}) - W_{jt}L_{jt} - c(I_{jt}),$$

where p is the output price, p^I is the price of one unit of capital, and $c(I_{jt})$ is the cost of investment.

"Optimizing out" labour

- At each point in time, labour will be chosen so as to maximize profits, conditional on capital. The first-order condition for labour implies:

$$\beta_l p K_{jt}^{\beta_k} L_{jt}^{\beta_l - 1} \exp(\beta_0 + \omega_{jt}) = W_{jt},$$

$$L_{jt} = \left(\frac{\beta_l p \exp(\beta_0 + \omega_{jt})}{W_{jt}} \right)^{\frac{1}{1-\beta_l}} K_{jt}^{\frac{\beta_k}{1-\beta_l}}.$$

Recall that we saw something similar earlier. Using this expression for labour in the profit function above, we can rewrite profits as

$$\begin{aligned} \Pi_t &= (1 - \beta_l) \beta_l^{\frac{\beta_l}{1-\beta_l}} (p \exp(\beta_0 + \omega_{jt}))^{\frac{1}{1-\beta_l}} (W_{jt})^{\frac{\beta_l}{\beta_l-1}} K_{jt}^{\frac{\beta_k}{1-\beta_l}} \\ &\quad - c(I_{jt}), \end{aligned}$$

or, in more reader-friendly notation,

$$\Pi_t = \varphi(W_{jt}, \omega_{jt}) K_{jt}^{\frac{\beta_k}{1-\beta_l}} - c(I_{jt}).$$

- You see how the labour variable has "disappeared" - replaced by the variables and parameters determining L_{jt} as implied by the first-order condition for labour. Using a notation more similar to that in OP, we might therefore write profits as

$$\Pi_t = \pi(k_{jt}, \omega_{jt}) - c(I_{jt})$$

where π_{jt} is sales minus labour costs, and $c(I_{jt})$ is the cost of investment.

The firm's investment demand

- It is assumed that the firm chooses investment and employment to maximize the present discounted value of current and expected future net revenues. We have already seen how labour is "optimized out" at each period, which means we can write the value of the firm as a function of capital and productivity only:

$$V(k_{jt}, \omega_{jt}) = \max_{I_t} E_t \sum_{s=t}^{\infty} \psi^{(s-t)} [\pi(k_{js}, \omega_{js}) - c(I_{js})],$$

where E_t denotes expectation given the information available in period t , and ψ is a discount factor.

The choice variable here is investment in period t .

- Note: the fact that labour is not visible in this equation does **not** mean labour is irrelevant. Labour is not visible here because we have implicitly replaced it by the variables and parameters **determining** labour as implied by the first-order condition. Indeed, estimating the coefficient on labour in the production function is a central objective in the analysis.
- Key for the OP approach is the firm's **investment**. In a model of the form outlined above, the firm is **forward looking** when choosing investment. Investment in period t will depend on
 - the existing capital stock; and
 - expectations about the future profitability of capital - i.e. expected future productivity.
- Because of the assumption that unobserved productivity is positively serially correlated, expected future productivity depends on current productivity. (recall: high productivity today \rightarrow high expected productivity tomorrow).
- OP hence write down an investment demand function of the following form:

$$I_{jt} = I_t(k_{jt}, \omega_{jt}).$$

- It is assumed that this function is **strictly increasing** in unobserved productivity - a firm with a high value of ω_{jt} will invest strictly more than a firm with a low value of ω_{jt} , conditional on k_{jt} .

Controlling for the endogeneity of input choice We are now ready to discuss the estimation strategy proposed by OP. Notice that this is motivated by the theory discussed above.

- **The key "trick" in OP.** Recall that investment is assumed to be a strictly monotonic in ω_{jt} .

This implies that the investment demand function

$$I_{jt} = I_{jt}(k_{jt}, \omega_{jt})$$

can be **inverted** so that productivity is expressed as a function of investment and capital:

$$\omega_{jt} = h_t(k_{jt}, I_{jt}).$$

- Intuitively, capital k_{jt} and investment I_{jt} "tell" us what ω_{jt} must be! This is the one-sentence summary of the OP approach.
- For example, suppose the investment demand function is as follows:

$$I_{jt} = \theta_0 + \theta_1 k_{jt} + \theta_2 \omega_{jt}.$$

Then it will also be true that

$$\omega_{jt} = \frac{I_{jt} - \theta_0 - \theta_1 k_{jt}}{\theta_2}$$

- Now return to the production function:

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + (\omega_{jt} + \eta_{jt}).$$

Recall that unobserved productivity ω_{jt} is a source of endogeneity bias.

- We now use $\omega_{jt} = h_t(k_{jt}, I_{jt})$ and rewrite the production function as

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + h_t(k_{jt}, I_{jt}) + \eta_{jt}.$$

By including the function $h_t(k_{jt}, I_{jt})$ as an additional term on the right-hand side, we have effectively "controlled" for unobserved productivity.

- Building on this, OP proposed a **two stage procedure** to estimate the parameters β_l and β_k . We will focus on the first stage - i.e. estimation of the labour coefficient.

- **First stage:** Define

$$\phi_t(k_{jt}, I_{jt}) = \beta_k k_{jt} + h_t(k_{jt}, I_{jt}),$$

and rewrite the production function

$$y_{jt} = \beta_k k_{jt} + \beta_l l_{jt} + (\omega_{jt} + \eta_{jt}).$$

as

$$y_{jt} = \beta_l l_{jt} + \phi_t(k_{jt}, I_{jt}) + \eta_{jt}.$$

- In general, the function ϕ_t is not linear. OP propose either approximating ϕ_t using a polynomial, e.g.

$$\phi_t(k_{jt}, I_{jt}) = \lambda_0 + \lambda_1 I_{jt} + \lambda_2 k_{jt} + \lambda_3 (I_{jt} \times k_{jt}) + \lambda_4 I_{jt}^2 + \lambda_5 k_{jt}^2,$$

or using kernel methods (nonparametric). In any case, what is clear now is that, provided we control for $\phi_t(k_{jt}, I_{jt})$, we may be able to identify the labour coefficient β_l in the first stage. Indeed, if we use the polynomial above, all we have to do is to estimate the following regression

$$y_{jt} = \lambda_0 + \beta_l l_{jt} + \lambda_1 I_{jt} + \lambda_2 k_{jt} + \lambda_3 (I_{jt} \times k_{jt}) + \lambda_4 I_{jt}^2 + \lambda_5 k_{jt}^2 + \eta_{jt}$$

using OLS.

- [EXAMPLE: Applying the first-stage OP procedure to the data underlying assignment 2. To be discussed in class.]
- The second stage, where we estimate the capital coefficient, is somewhat more complicated than the first stage, and you may skip this part if you wish. For those interested, I outline the second stage in the appendix.

4.3.2. Discussion

Whilst theoretically elegant, the OP approach won't always work. Theoretical reasons as to why the OP estimator may not work are carefully discussed in the paper by Akerberg et al. (2006). Because the details of this discussion, however, are quite technical and beyond the scope of the course, I simply list the main points here.

- There may be more than one productivity factor. Recall the OP model assumes unobserved productivity is equal to ω_{jt} . However, if there are **two** unobserved productivity factors, say ω_{jt}^1 and ω_{jt}^2 , the OP approach won't work, because there is no way of fully characterizing unobserved productivity by investment and capital. Recall we said that capital k_{jt} and investment I_{jt} "tell" us what ω_{jt} must be - but they cannot tell us what ω_{jt}^1 and ω_{jt}^2 are separately, if they are both relevant.
- Zero investment levels potentially problematic. Investment needs to be a strictly monotonic function of unobserved productivity. The presence of lots of zero investments in the data strongly indicates that this is not the case - it's unrealistic to assume that all firms that invest nothing have precisely the same level of unobserved productivity (conditional on capital). Again, in this case, k_{jt} and investment I_{jt} will not tell us what ω_{jt} is - as zero investment may be associated with different values of ω_{jt} .
- Labour really flexible? The OP approach just described is really **only** appropriate if labour is a flexible input. If not, e.g. because firms can't easily hire and fire workers from one day to another, then OP approach won't work.
- Awkward assumptions. Wages need to vary across firms, and be serially uncorrelated; yet there must be **no** variation across firms in the **cost of capital**. Do you really believe this?

Illustration: Average and marginal cost in the presence of fixed costs

$$\text{Total cost} = q^{1.5} + 5$$

Note: fixed cost = 5

$$\text{Average cost} = (q^{1.5} + 5)/q$$

$$\text{Marginal cost} = 1.5 \cdot q^{0.5}$$

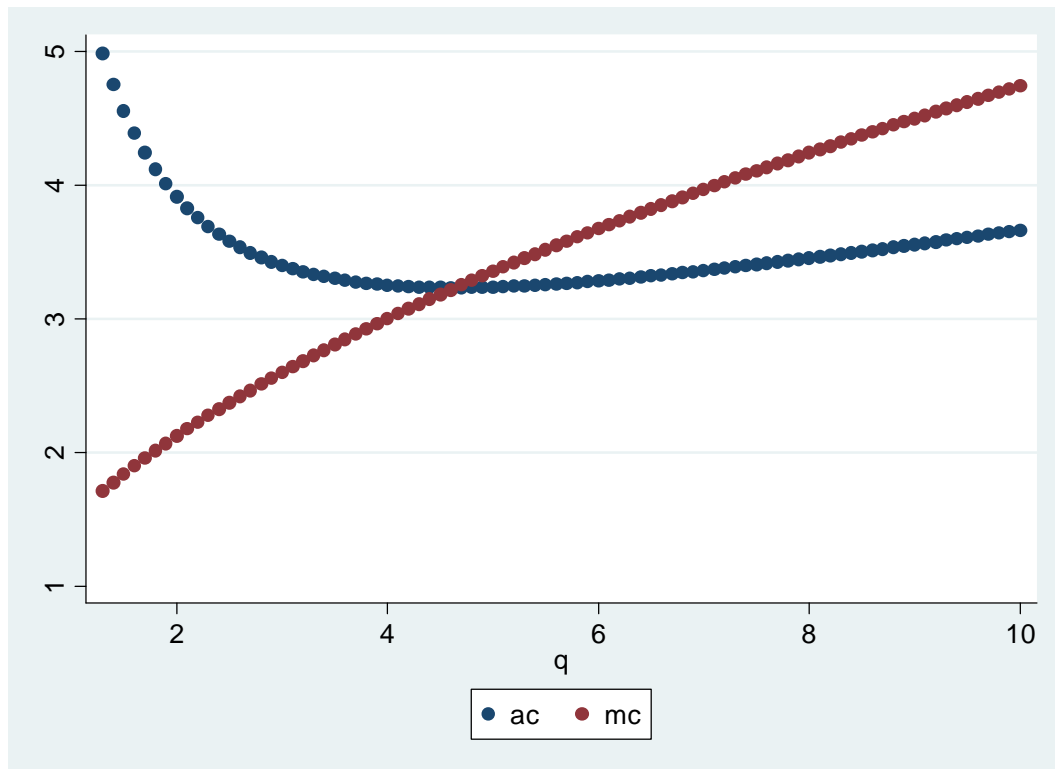


Illustration: CR4 and population in the food retailer data (assignment 1)

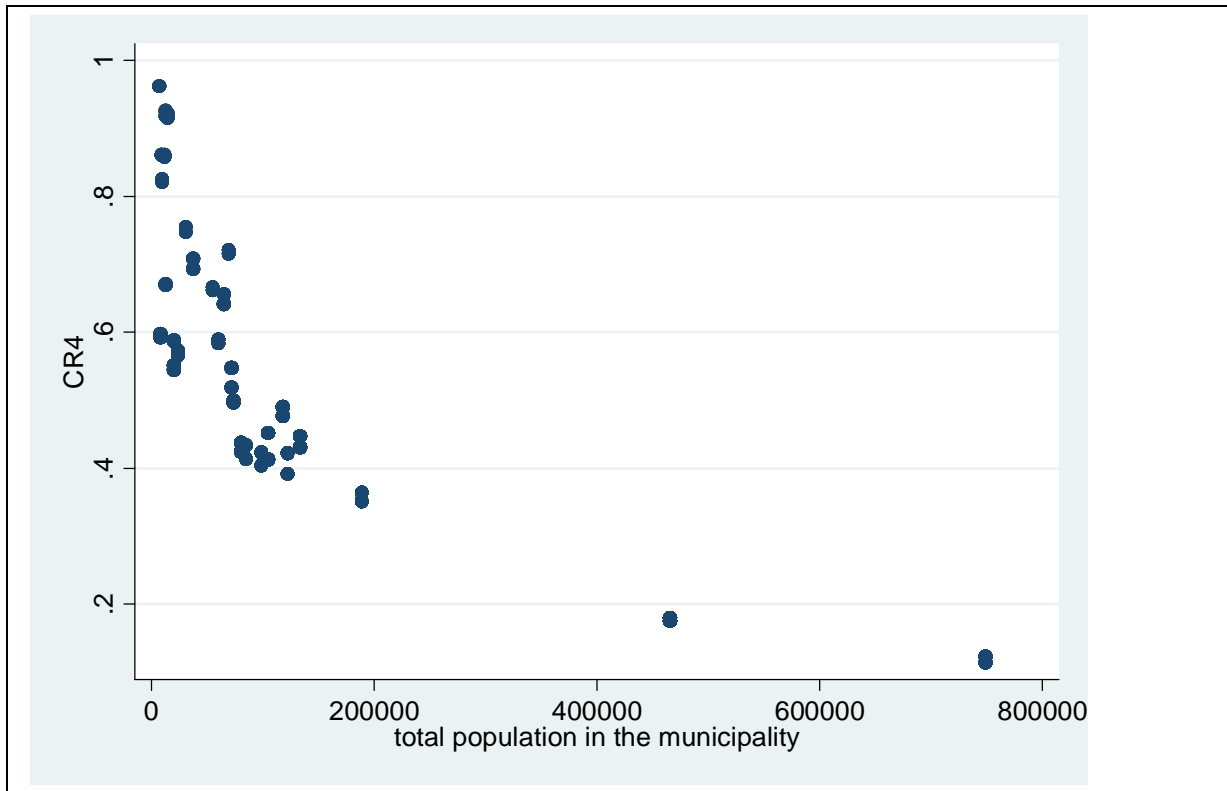


Illustration: log output per employee (y/l) vs. log capital per employee (k/l) in US manufacturing firms (assignment 2)

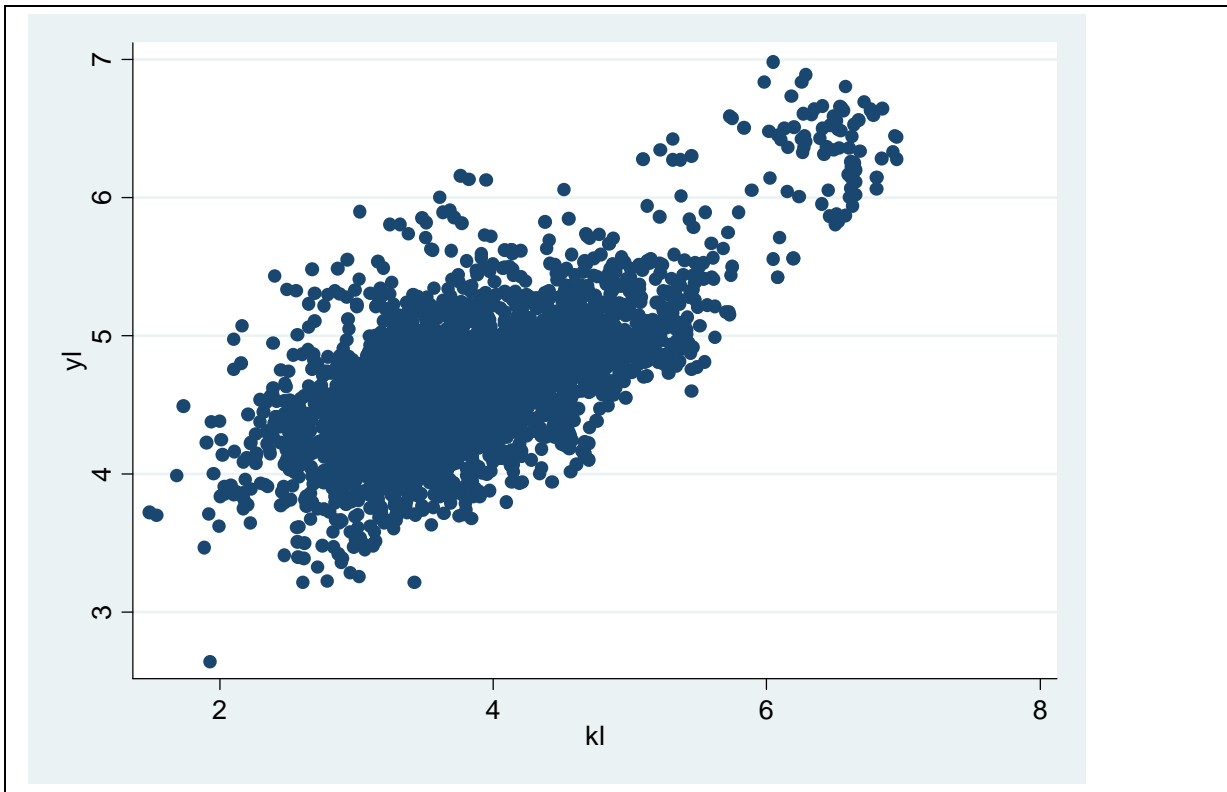


Illustration: OP with linear terms only, interacted with time

```
. xi: reg y i.year*k i.year|ik n , robust cluster(id)
i.year      _Iyear_1982-1989      (naturally coded; _Iyear_1982 omitted)
i.year*k    _IyeaXk_#              (coded as above)
i.year|ik   _IyeaXik_#            (coded as above)
```

Linear regression

Number of obs = 3563
 F(21, 508) = 1127.34
 Prob > F = 0.0000
 R-squared = 0.9692
 Root MSE = .35206

(Std. Err. adjusted for 509 clusters in id)

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
_Iyear_1983	-.0117025	.0397574	-0.29	0.769	-.0898116	.0664066
_Iyear_1984	-.0003504	.0386859	-0.01	0.993	-.0763545	.0756537
_Iyear_1985	.0087549	.0338052	0.26	0.796	-.0576604	.0751701
_Iyear_1986	-.0039311	.0271325	-0.14	0.885	-.0572368	.0493746
_Iyear_1987	-.0064952	.0213817	-0.30	0.761	-.0485027	.0355122
_Iyear_1988	(dropped)					
_Iyear_1989	-.0085923	.0218335	-0.39	0.694	-.0514875	.0343028
k	.4305507	.0278018	15.49	0.000	.37593	.4851714
_IyeaXk_1983	.0003229	.0039407	0.08	0.935	-.0074191	.0080649
_IyeaXk_1984	(dropped)					
_IyeaXk_1985	-.0045234	.0034495	-1.31	0.190	-.0113004	.0022536
_IyeaXk_1986	-.0071593	.0049993	-1.43	0.153	-.0169812	.0026626
_IyeaXk_1987	.0011513	.0052254	0.22	0.826	-.0091146	.0114173
_IyeaXk_1988	.0069001	.0056709	1.22	0.224	-.0042412	.0180414
_IyeaXk_1989	.0056234	.0059086	0.95	0.342	-.0059849	.0172316
ik	-.1679462	.1046535	-1.60	0.109	-.3735531	.0376607
_IyeaXi~1983	.2370482	.1352143	1.75	0.080	-.0285999	.5026963
_IyeaXi~1984	.2041958	.1285209	1.59	0.113	-.048302	.4566937
_IyeaXi~1985	(dropped)					
_IyeaXi~1986	.2056813	.1355581	1.52	0.130	-.0606422	.4720048
_IyeaXi~1987	.1325166	.1263455	1.05	0.295	-.1157073	.3807406
_IyeaXi~1988	.1283877	.148174	0.87	0.387	-.1627217	.419497
_IyeaXi~1989	.124821	.1406473	0.89	0.375	-.1515009	.401143
n	.5609091	.0303342	18.49	0.000	.5013132	.620505
_cons	2.996726	.0999766	29.97	0.000	2.800308	3.193145