Firms and Development

Lecture notes for Development Economics I
(option course, MPhil in Economics)

Simon Quinn*

Michaelmas 2022

Outline of these lectures

These notes provide some background for four lectures that I will be giving this year for the Development Economics I option course. I will use slides for the lectures, and I will make these slides available online *after* the lectures. It is likely that there will be some things in these notes that we do not have time to cover in class, and we *may* cover some things in class that are not covered in these notes.

I have suggested a list of references for each lecture; you should read the ones that are starred ('⋆'). Among the starred readings, you should pay more attention to the readings discussed in these notes; the other starred readings are required as background to the core concepts we cover. You are not expected to read the non-starred references. Though we will focus in class on the most important issues, please consider all of the lectures, all of the notes and all of the starred readings to be potentially relevant for the exam.

*Department of Economics, Centre for the Study of African Economies and St Antony's College, University of Oxford: simon.quinn@economics.ox.ac.uk. Without implicating them in any shortcomings, I would like to thank Amrit Amirapu, Rachel Cassidy, Markus Eberhardt, Marcel Fafchamps, Sanghmitra Gautam, Anne Karing, Muhammad Meki, Justin Sandefur, Daniela Scur, Francis Teal, Jacek Witkowski and Andrew Zeitlin.

. . . but do firms matter?

Why study firms in developing countries? I think there are several important reasons. First, *firms matter for policy*. Many aspects of government policy — including, in particular, job creation — depend critically upon firm performance. The Project Overview for the DFID's Private Enterprise Development in Low-Income Countries Research Initiative makes this point very directly:

> It is impossible for large numbers of people to be lifted out of poverty without sustained growth, and impossible to have sustained growth without a vibrant private sector. . . sustained growth which leads to rising income and creation of jobs is impossible to achieve without growth in productivity.

Second, *firms are central to the welfare of most people in developing economies*. In these lectures, we will think of the concept of a firm very broadly — as encompassing not just large corporations, but also small household-based enterprises (including farms). In this sense, the study of firms in developing economies is the study of how large numbers of poor people make critical decisions about how to organise their time and resources, and the contraints that they face in doing so. This is one important reason that I think the study of firms is important: when we study firms in developing economies, we are forced to think about the poor as *active decision-makers in their own futures*. This is a view of poverty that can sometimes be overlooked in other areas of development economies — for example, it is not uncommon in development economics for poor households to be portrayed implicitly as mere passive recipients of donor and government aid programs.

Third, *firms are fundamental to any process of structural transformation*. We often think about the process of long-term development as one of shifting employment out of agriculture and into manufacturing and services; if this is so, then understanding the structure and producitivity of firms — both agricultural and otherwise — is critical for appreciating how and why different economies grow and transform in different ways at different times.

Our approach in these lectures

These lectures do *not* attempt to summarise generally 'what we have learned about firms' in developing countries; in my view, that is inherently a moving target, and a topic that can be addressed adequately by literature reviews rather than lectures. Instead, these lectures aim to provide an overview of some *issues*, *models* and *empirical methods* that assist in studying firms and development. The issues are framed in terms of the following four lecture titles; in each lecture, we will consider both a theoretical model and an empirical method.

Michaelmas			
Week 1	Lecture 1	**The Firm and Accumulation**	Monday, 2pm – 4pm Lecture Theatre
Week 1	Lecture 2	**The Firm and Production**	Wednesday, 2pm – 4pm Seminar Room A
Week 2	Lecture 3	**The Firm and Technology Adoption**	Monday, 2pm – 4pm Lecture Theatre
Week 2	Lecture 4	**Firm Size**	Wednesday, 2pm – 4pm Seminar Room A

1 Lecture 1: The Firm and Accumulation

References:

- ⋆ Bernhardt, A., Field, E., Pande, R., and Rigol, N. (2019): "Household Matters: Revisiting the Returns to Capital among Female Micro-entrepreneurs," *American Economic Review: Insights*, 1(2), 141–160.
- ⋆ de Mel, S., McKenzie, D., and Woodruff, C. (2012): "One-Time Transfers of Cash or Capital Have Long-Lasting Effects on Microenterprises in Sri Lanka," *Science*, 335, 962–966.
- ⋆ Fafchamps, M., McKenzie, D., Quinn, S., and Woodruff, C. (2014): "Microenterprise Growth and the Flypaper Effect: Evidence from a Randomized Experiment in Ghana," *Journal of Development Economics*, 106, 211-226.
- Acemoglu, D. (2009): *Introduction to Modern Economic Growth.* Princeton University Press, chapter 6.
- Adda, J., and Cooper, R. (2003): *Dynamic Economics.* The MIT Press, chapter 2.
- Athey, S. and Imbens, G.W. (2016): "The Econometrics of Randomized Experiments," *Working paper: arXiv:1607.00698v1*.
- Bandiera, O., Barankay, I., and Rasul, I. (2011): "Field Experiments with Firms," *Journal of Economic Perspectives*, 25(3), 63–82.
- Banerjee, A., Fischer, G., Karlan, D., Lowe, M., and Roth, B. (2022): "Does the Invisible Hand Efficiently Guide Entry and Exit? Evidence from a Vegetable Market Experiment in India," *Working paper*.
- Bruhn, M., and McKenzie, D. (2009): "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1(4), 200–232.
- Deaton, A. and Cartwright, N. (2018): "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine*, 210, 2–21.
- de Mel, S., McKenzie, D., and Woodruff, C. (2008): "Returns to Capital in Microenterprises: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 123(4), 1329–1372.
- Field, E., Pande, R., Papp, J., and Rigol, N. (2013): "Does the Classic Microfinance Model Discourage Entrepreneurship Among the Poor? Experimental Evidence from India," *American Economic Review*, 103(6), 2196 – 2226.
- Słoczyński, T. (2022): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, 104(3): 501–509.

We begin by considering the problem of *capital accumulation*. Every firm needs to decide how much to invest, and when to do so — so the problem of capital accumulation is a general one, which will form a useful foundation for the rest of this module. But the problem of capital accumulation also has *specific* relevance for thinking about the way that 'microenterprises' divide their resources between production and consumption. In this lecture, we will:

(i) Use a growth model to think about investment in a microenterprise;

(ii) Discuss a 'balanced randomisation' method for testing the key implications of the growth model, and

(iii) Consider empirical results from microenterprises in urban Ghana.

All of this broadly follows Fafchamps *et al*; we will then conclude by briefly considering other recent research on the returns to capital in microenterprises.

Introduction: 'Own-account workers'

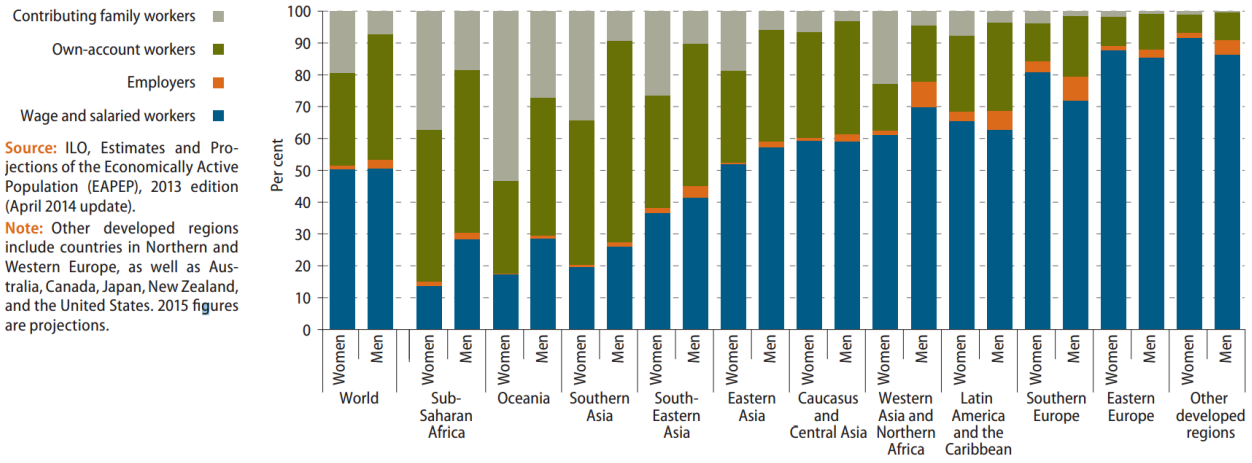
It is tempting to think of 'firms' as referring to large and formal organisations — possibly with a clear management structure or formal legal recognition. However, the concept need not be so narrow, for the models and methods that we will use can apply readily to a much broader set of economic activities. In this lecture, we begin our discussion of firms and development by considering enterprises that are generally very small and often not legally formalised. These are what the literature often terms 'microenterprises' — or, in the terminology of the International Labour Organisation, 'own-account workers':

> "those workers who, working on their own account or with one or more partners, hold the type of job defined as a self- employed job, and have not engaged on a continuous basis any employees to work for them during the reference period."[1]

Figure 1.1 shows the proportion of workers that are 'own-account workers', across a variety of developing countries; the figure is reproduced from '*World's Women 2015*, a publication of the United Nations Statistics Division. Even from this simple graph, it is clear (i) that the microenterprise is a very important form of organisation for economic activity, and (ii) that it is an important form of activity for women (as, for example, many narratives of microfinance attest); however, it also shows (iii) that a relatively large proportion of *men* also rely upon this form. In this lecture, we will consider both theoretical and empirical results on the way that such microenterprises use capital.

[1] See, for example, http://unstats.un.org/unsd/mdg/Metadata.aspx?IndicatorId=0&SeriesId=773. The ILO considers every own-account worker and every unpaid family worker to be in 'vulnerable employment'. I leave it to you to consider whether this is either a reasonable generalisation of microenterprises or a sensible approach for analysing income vulnerability.

Figure 1.1: **"Distribution of employment by status in employment, by sex and region, 2015" (Figure 4.10, *The World's Women 2015*, published by the UN Statistics Division)**



1.1 Theoretical model: A Ramsey framework in discrete time

When economists talk about 'growth theory', they are usually referring to a topic in macroeconomics — the theory of how national economies increase production over time, and of why some economies may increase production faster than others. But the basic problem of economic growth is a problem of intertemporal accumulation: *what share of resources should be devoted to production, and what share should be consumed?* This is a question that applies to microenterprises as much as it applies to national governments. In this section, we apply a standard growth framework to think about the intertemporal decision facing microenterprises.

Fafchamps *et al* present a general model framework, from which the authors derive general testable predictions; however, for the purposes of a clear discussion, we will consider a slightly modified model with specific functional forms.

Assumption 1 (Timing) *The firm takes decisions at discrete time periods, and holds wealth between periods only as capital (k_t). In period t, the firm begins with capital k_t and takes c_t for household consumption. Production then occurs, using the remaining capital $k_t - c_t$.*

Assumption 2 (Production) *Production is a function of remaining capital and the firm's time-invariant 'ability' (θ):*

$$k_{t+1} = \pi(k_t - c_t, \theta) + k_t - c_t \tag{1.1}$$
$$= f(k_t - c_t, \theta). \tag{1.2}$$

*The production function (and, therefore, the function $f(\cdot)$) is real-valued, differentiable, strictly increasing and strictly concave; further, we assume that $\lim_{x\to\infty} \partial\pi(x,\theta)/\partial x = 0 \ \forall \ \theta$ and $\lim_{x\to 0} \partial\pi(x,\theta)/\partial x = \infty \ \forall \ \theta$. (That is, we assume that $\pi(\cdot)$ satisfies the **Inada conditions**.) Additionally, we assume that entrepreneurs with higher ability have (i) higher production for given capital and (ii) higher marginal returns to capital:*

$$\frac{\partial\pi(x,\theta)}{\partial\theta} > 0; \ \frac{\partial^2\pi(x,\theta)}{\partial x\partial\theta} > 0. \tag{1.3}$$

Assumption 3 (Preferences) *The firm has time-separable preferences for consumption using exponential discounting with discount factor δ:*

$$U_0 = \sum_{t=0}^{\infty} \delta^t \cdot u(c_t). \tag{1.4}$$

(The assumption of time-separable exponential discounting is a very strong one. Fafchamps *et al* consider an extension to the model in which the firm uses hyperbolic discounting — under a so-called 'beta-delta' framework — but we will not discuss that case in this lecture.)

We can therefore write the firm's optimisation problem as:

$$\max_{\{c_t\}_{t=0}^{\infty}, \{k_t\}_{t=1}^{\infty}} \sum_{t=0}^{\infty} \delta^t \cdot u(c_t) \text{ subject to}$$
$$k_{t+1} = f(k_t - c_t, \theta); \tag{1.5}$$
$$c_t > 0; \tag{1.6}$$
$$k_t > 0. \tag{1.7}$$

As Acemoglu (2009, page 185) explains, this optimisation problem "...corresponds to a *sequence problem*; it involves choosing an infinite sequence $[\{k_t\}_{t=1}^{\infty}]$. Sequence problems sometimes have nice features, but their solutions are often difficult to characterise both analytically and numerically."

1.1.1 First-order conditions and comparative statics

We will shortly make some very specific assumptions about the form of $u(\cdot)$ and $f(\cdot)$. But before we do that, we can characterise the first-order conditions generally. We can write the following Lagrangian:

$$\mathcal{L}(c_0, c_1, \ldots, k_0, k_1, \ldots; \lambda_0, \lambda_1, \ldots) = \sum_{t=0}^{\infty} \left(\delta^t \cdot u(c_t) + \lambda_t \cdot [k_{t+1} - f(k_t - c_t, \theta)] \right). \tag{1.8}$$

Differentiating,

$$\frac{\partial \mathcal{L}(\cdot)}{\partial c_t} = \delta^t \cdot u'(c_t) + \lambda_t \cdot f'(k_t - c_t, \theta) \tag{1.9}$$

$$\frac{\partial \mathcal{L}(\cdot)}{\partial k_t} = \lambda_{t-1} - \lambda_t \cdot f'(k_t - c_t, \theta), \tag{1.10}$$

where $f'(x, \theta)$ refers to $\partial f(x, \theta)/\partial x$. You should confirm that, by setting these partial derivates to zero and rearranging, we can obtain the following Euler equation:

$$\frac{u'(c_t)}{\delta \cdot u'(c_{t+1})} = f'(k_t - c_t, \theta). \tag{1.11}$$

If there exists a steady state (such that $c_t = c^* \,\forall\, t$ and $k_t = k^* \,\forall\, t$), it follows straightforwardly from the Euler equation that we can characterise that state by the following conditions:

$$f'\left(k^* - c^*, \theta\right) = \frac{1}{\delta} \Leftrightarrow \Delta c_t = 0 \tag{1.12}$$

$$\text{and } k^* = f\left(k^* - c^*, \theta\right) \Leftrightarrow \Delta k_t = 0. \tag{1.13}$$

Figure 1.2 illustrates; this style of diagram should be familiar from growth theory.

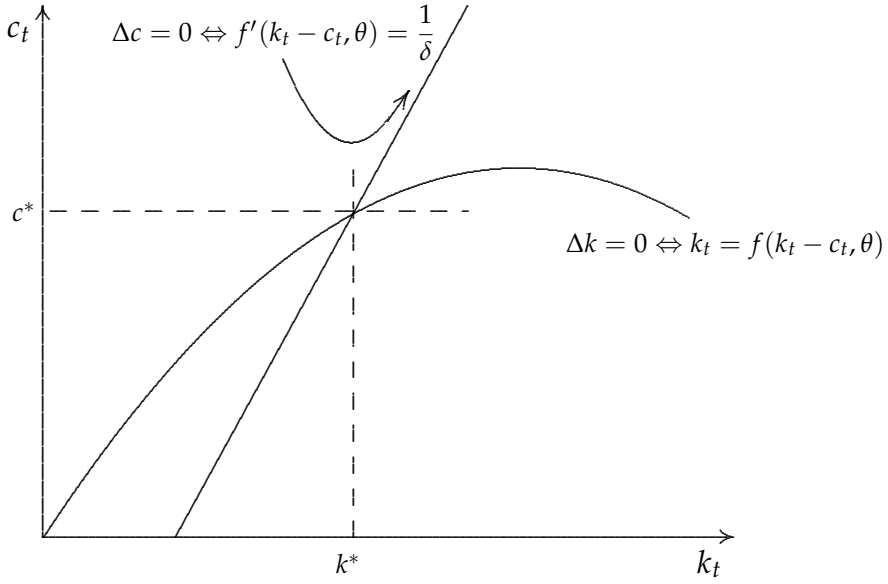
You should confirm that:

(i) More patient firms have more capital in the steady state (k^*) and higher retained capital ($k^* - c^*$);

(ii) The same result obtains for firms with higher ability (*i.e.* firms with higher θ).

(Can we say anything about $dc^*/d\delta$? What about $dc^*/d\theta$?)

Figure 1.2: **Phase diagram: A Ramsey model in discrete time**



1.1.2 A Bellman equation and a policy function

The comparative statics just outlined are useful. However, we are particularly interested in thinking about the response of the firm to a capital shock (for example, a researcher arriving and providing a 'capital drop'). That is, we are interested not merely in the characteristics of the steady state but also in the *transition* to that steady state. We can solve this analytically with some convenient specific assumptions for functional form.

Assumption 4 (Logarithmic utility and Cobb-Douglas production)

$$u(c_t) = \ln c_t; \tag{1.14}$$
$$f(k_t - c_t, \theta) = \theta(k_t - c_t)^{\alpha}. \tag{1.15}$$

Using the earlier results, we can now characterise two curves in (k_t, c_t) space:

$$\Delta c = 0 \Leftrightarrow c_t = k_t - (\alpha\theta\delta)^{\frac{1}{1-\alpha}}; \tag{1.16}$$
$$\Delta k = 0 \Leftrightarrow c_t = k_t - \left(\frac{k_t}{\theta}\right)^{\frac{1}{\alpha}}. \tag{1.17}$$

In order to think about the consequences of a capital shock, we need to do much more than merely characterise the steady state; we need to solve for the ***policy function*** — a function that maps from any level of the capital stock into optimal consumption. Acemoglu (2009, p.185) explains: "Intuitively, ... instead of explicitly choosing the sequence $[\{k_t\}_{t=1}^{\infty}]$, we choose a *policy*, which determines what the [control variable $c(t)$] should be

for a given value of the [state variable $k(t)$]." In this case, we can solve for this function using a Bellman equation.

We can write the firm's optimisation problem — without further loss of generality — with the following Bellman equation:

$$V(k_t) = \max_{c_t \in (0,k_t]} \{\ln c_t + \delta \cdot V(k_{t+1})\} \tag{1.18}$$

$$= \max_{c_t \in (0,k_t]} \left\{\ln c_t + \delta \cdot V\left(\theta \left(k_t - c_t\right)^{\alpha}\right)\right\}. \tag{1.19}$$

You should recall:

- That $V(k)$ is called a '***value function***'; it tells us the *value* to the firm of holding capital k, assuming that the firm will use k optimally. It is directly analogous to the *indirect utility function* in consumer theory, which tells us the value to the consumer of a given income and price vector, assuming that the consumer will spend their income optimally given prices.
- That we can refer to c as the '***control variable***' — the variable chosen by the firm.
- That we can refer to k as the '***state variable***'; as Adda and Cooper (2003, p.16) explain, "The state completely summarises all information from the past that is needed for the forward-looking optimisation problem."
- That we can refer to equation 1.1 (on page 7) as the '***transition equation***'.
- That the Bellman equation is known as a '***functional equation***'; as Adda and Cooper (2003, p.17) put it, "...the unknown in the Bellman equation *is the value function itself*: the idea is to find a function [$V(k)$] that satisfies this condition for all k" (emphasis added). Thus, as Acemoglu (2009, p.185) explains, "The basic idea of dynamic programming is to turn the sequence problem into a *functional equation*; that is, to transform the problem into one of finding a function rather than a sequence" (emphasis in original).

You should verify that this particular Bellman equation can be satisfied by a value function of the form

$$V(k_t) = \frac{\ln k_t}{1 - \delta\alpha} + C, \tag{1.20}$$

where C represents a constant, and that this implies the following policy function:

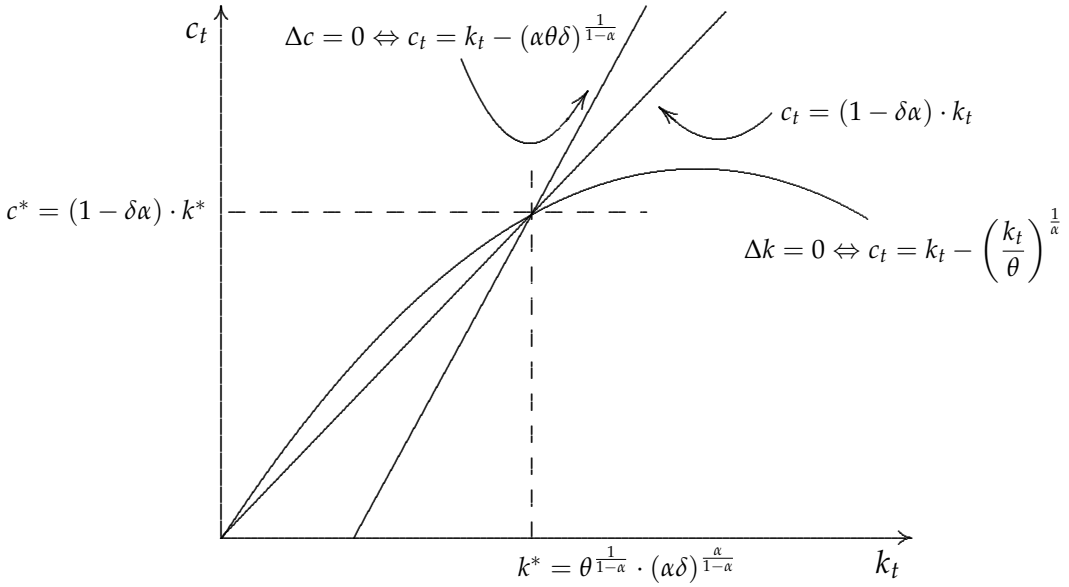
$$c_t(k_t) = (1 - \delta\alpha) \cdot k_t. \tag{1.21}$$

(I show how this can be verified in the appendix to this chapter. You may be concerned about whether this solution is *unique*; it is: see Adda and Cooper (2003, Chapter 2).)

This, of course, is the famous 'constant proportion' result from, say, cake-eating problems under logarithmic utility. Note that the agent consumes a higher proportion of capital if (i) the agent is less patient, and/or (ii) the returns to capital are lower.

Figure 1.3 illustrates; it extends Figure 1.2 by adding (i) the policy function and (ii) specific functional forms.

Figure 1.3: **Phase diagram: Log utility with Cobb-Douglas production**



1.1.3 Introducing a financial asset

Fafchamps *et al* also consider the role of a financial asset, w_t, with an annual interest return of r; this could represent, for example, the value of money held with a bank. This is particularly useful for thinking about how and why the effects of 'cash shocks' may differ from 'in-kind shocks' (as we will discuss shortly). However, this is a difficult case to consider with an analytical solution; effectively, the introduction of a financial asset creates a *second state variable*, and we would need to solve for a *bivariate* value function $V(k_t, w_t)$. My sense is that it would be very difficult – if not impossible – to solve analytically a value function for this case. Of course, merely because we may not be able to find an analytical ('closed-form') solution for a value function does not mean that the problem is intractable. Functional equations can usually be solved by numerical methods, and this may be the only way to explore all of the possibilities in this 'financial asset' case.[2]

[2] The paper by Bond and Söderbom (2005), listed on the reference list for the next lecture, provides an interesting illustration of how numerical methods can be used to solve for a value function by iteration.

1.1.4 Testable implications

Fafchamps *et al* study the consequences of providing gifts to Ghanaian microenterprises — either in the form of cash or in-kind. In this way, the authors follow the seminal work of de Mel *et al* (2008), who studied the consequences of cash and in-kind transfers to microenterprises in Sri Lanka.

The present theoretical model implies that, for any microenterprise below its steady state, *both* cash and in-kind transfers should increase profits and capital. Conversely, transfers may have different effects for firms having reached a steady state: for example, a firm having reached k^* should merely increase savings (including, if necessary, liquidating any in-kind transfer). See page 6 of Fafchamps *et al* for a more detailed discussion of these predictions.

Fafchamps *et al* therefore estimate models of the following form:

$$\pi_{i,t+s} = \beta_1 \cdot M_{it} + \beta_2 \cdot E_{it} + u_{i,t+s} \tag{1.22}$$

$$k_{i,t+s} = \alpha_1 \cdot M_{it} + \alpha_2 \cdot E_{it} + v_{i,t+s}, \tag{1.23}$$

where π refers to profits, k refers to capital, and M_{it} and E_{it} are dummy variables recording whether firm i has received respectively either a cash or in-kind treatment by period t.

As the authors explain (pp.12–13),

> The [Ramsey] models predict $\alpha_1 = \alpha_2 > 0$ and $\beta_1 = \beta_2 > 0$ if the firm was below its steady state at the time of the treatment. They also predict $\alpha_1 = \beta_1 = 0$ if the firm had already reached its equilibrium size at time t... Because the in-kind treatment is not immediately fungible, these models also predict $\alpha_2 > 0$ and $\beta_2 > 0$ for a small time from treatment s, but eventually $\alpha_2 = \beta_2 = 0$ for s large enough, as k returns to its steady state from above.

But, as we shall see, things are rarely so straightforward...

1.2 Empirical method: A Randomised Controlled Trial ('RCT')

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested...

> Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order... that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.
>
> R.A. Fisher, *The Design of Experiments*, 1935

Whatever else Randomised Controlled Trials may be, they are certainly not new. But, as every economist knows, randomisation has grown in the past 15 years to assume a new prominence — both in development economics and beyond. The justification for randomising in this general context is straightforward, and is explained concisely by Karlan and Zinman (2011, page 1283, footnote 14):

> ...a concern with nonrandomized studies is that those who get credit (because of relatively high demand or high supply) get it because they have businesses or other investment options with relatively high returns, and those who seem similar on observable characteristics but do not choose to borrow simply do not have the same set of business or investment options. A methodology that does not account for, or remove, this sort of underlying correlation may mistakenly attribute causal impacts to microcredit use/access that are actually due to underlying differences between borrowers and nonborrowers that have nothing to do with microcredit per se. The same concern holds for other characteristics that are difficult to observe (and hence control for) that might be correlated with both outcomes (e.g. business success) and borrowing or lending decisions.

This clearly does not imply, of course, that the results of RCTs should be lexicographically preferred to results from other empirical methods; indeed, there remains an important ongoing debate about how and when to rely upon the results from randomised evaluations.[3] But the RCT is a very useful tool for thinking about the incentives and constraints faced by microenterprises, as Bandiera, Barankay and Rasul (2011) explain.

Randomised Controlled Trials: How? We sometimes seem to spend much more time extolling the *general virtues* of randomisation rather than discussing *specific methods* by which it takes place. Bruhn and McKenzie (2009) is an important exception to that principle: this paper surveys various methods of randomisation, and makes recommendations for methods for 'balancing' a randomisation. Those methods can be summarised briefly

[3] See, for example, Deaton and Cartwright (2018). One important aspect of this debate, for example, may be a trend away from 'mere program evaluation' towards the use of randomisation in order to test specific economic models and mechanisms.

in the context of the estimation used in Fafchamps *et al*.

We noted earlier that Fafchamps *et al* were interested in testing the effect of cash and in-kind transfers on microenterprise profits, implying the relationship in equation 1.22. Fafchamps *et al* followed 792 firms, of which 198 received in-kind transfers, 198 received cash and 396 were left as a 'control group'. Had the authors merely assigned these groups randomly from the entire pool of 792 firms, the estimates of β_1 and β_2 (and, of course, of α_1 and α_2) *would*, in principle, be unbiased: after all, the 'treatment' would still be orthogonal to other firm characteristics.

However, Bruhn and McKenzie (2009) show that, in many cases, researchers can do better than this. Suppose that, by luck, those firms assigned to the treatment are larger — or more likely to be owned by men, or more likely to have reported vulnerability to external pressures, and so forth. *A priori*, we would still describe the randomisation as producing unbiased estimates — however, in practise, we may be concerned that the treatment is *proxying* for other important characteristics. We could, of course, simply include baseline characteristics in the estimation (or estimate using fixed effects), but this may not necessarily be the most efficient approach.

Instead, Fafchamps *et al* — following Bruhn and McKenzie — first *stratified* firms on the basis of several important characteristics (specifically, gender, sector, capital stock and a measure of 'high capture' vulnerability). Specifically, the authors formed *matched quadruplets*: "[w]ithin the quadruplet one firm was then randomly chosen to receive the cash treatment, one to receive the in-kind treatment, and two to be control firms" (Fafchamps *et al*, p.17). Therefore, rather than using equation 1.22 directly, the authors estimated using:

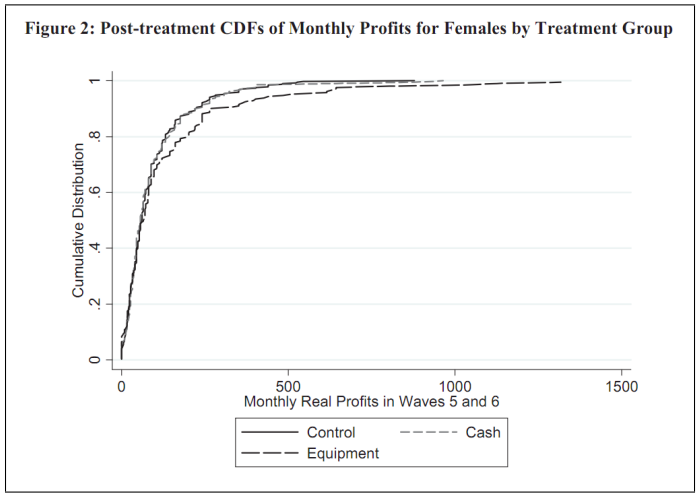
$$\pi_{it} = \beta_1 \cdot M_{it} + \beta_2 \cdot E_{it} + \sum_t \delta_t \cdot D_{it} + \sum_{g=1}^{G} \gamma_g \cdot S_{ig} + \varepsilon_{it}, \tag{1.24}$$

where S_{ig} is a dummy variable for each matched quadruplet. In effect, this approach ensures that comparison of treated and control occurs *within* each quadruplet, rather than simply across the sample as a whole. In this way, the assignment to treatment and control can be both randomised *and balanced*.[4]

[4] For a detailed discussion of econometric issues surrounding the design and analysis of randomised experiments, I strongly recommend the recent work of Athey and Imbens (2016). The details presented in this paper go beyond the scope of this course — but I strongly recommend their discussion if you are interested in learning more, and particularly if you are considering running a randomised experiment as part of your own research. More generally, if you are interested in using a regression structure like equation 1.24 in a context where the Average Treatment on the Treated is likely to differ from the Average Treatment on the Untreated, be sure to see the recent work of Słoczyński (2020).

1.3 Empirical results: Returns to capital in Ghanaian microenterprises

(i) Interpret the following graph. How does it relate to the empirical results?

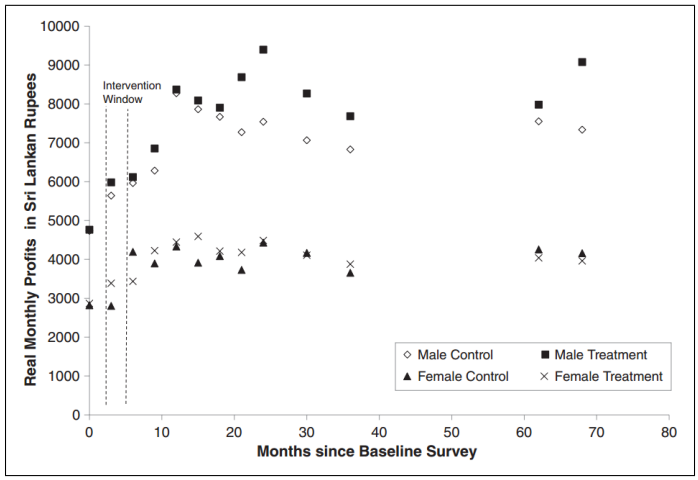


(ii) The authors refer to 'asset integration'. What do they mean by this?

(iii) Where did the grants go? Is there a coherent explanation for the authors' results on this?

(iv) What policy implications — if any — should follow from these results?

(v) Interpret the following graph from de Mel *et al* (2012). How might these results challenge a simple Ramsey framework for thinking about microenterprise growth?



(vi) Consider the recent work of Bernhardt *et al* (2019). How should this change our interpretation of the results in Fafchamps *et al* (2014)?

Appendix to Lecture 1: Verifying the value function

Assume that the form of the value function is correct. Then, for some given k, we can write the firm's utility function as:

$$U(c;k) = \ln c + \delta \cdot \left\{ \frac{\ln\left[\theta(k-c)^{\alpha}\right]}{1-\delta\alpha} + C \right\}. \tag{1.25}$$

So this is maximised by:

$$\left.\frac{\partial U(c;k)}{\partial c}\right|_{c=c^*(k)} = \frac{1}{c} - \frac{\delta\alpha}{(1-\delta\alpha)\cdot(k-c)} = 0 \tag{1.26}$$

$$\therefore \delta\alpha \cdot c^*(k) = (1-\delta\alpha)\cdot(k-c^*(k)) \tag{1.27}$$

$$\therefore c^*(k) = (1-\delta\alpha)\cdot k. \tag{1.28}$$

So the optimal policy function will be $c_t(k_t) = (1-\delta\alpha)\cdot k$, *if* the form of the value function is correct. We can now check this:

$$V(k) = \ln c^*(k) + \delta \cdot \left\{ \frac{\ln\left[\theta(k-c^*(k))^{\alpha}\right]}{1-\delta\alpha} + C_1 \right\} \tag{1.29}$$

$$= \ln\left[(1-\delta\alpha)\cdot k\right] + \delta \cdot \left\{ \frac{\ln\left[\theta(\delta\alpha\cdot k)^{\alpha}\right]}{1-\delta\alpha} + C_1 \right\} \tag{1.30}$$

$$= \ln k + \left(\frac{\delta\alpha}{1-\delta\alpha}\right)\cdot \ln k + C_2 \tag{1.31}$$

$$= \frac{\ln k}{1-\delta\alpha} + C_2. \tag{1.32}$$

Thus the form of the value function is correct, and we have derived the optimal policy function.

2 Lecture 2: The Firm and Production

References:

- ⋆ Ackerberg, D., Caves, K. and Frazer, G. (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83(6), 2411–2451.
- ⋆ Bond, S., and Söderbom, M. (2005): "Adjustment Costs and the Identification of Cobb Douglas Production Functions," *IFS Working Paper WP05/04*.
  I encourage you to read all of this paper; however, only Sections 1–3 are required.
- ⋆ Eberhardt, M., and Helmers, C. (2019): "Untested Assumptions and Data Slicing: A Critical Review of Firm-Level Production Function Estimators," *working paper*.
  I encourage you to read all of this paper, too; however, you are required to read only Sections 1 and 2.
- ⋆ Söderbom, M., and Teal, F. (2004): "Size and Efficiency in African Manufacturing Firms: Evidence from Firm-Level Panel Data," *Journal of Development Economics*, 73(1), 369–394.
- Arellano, M., and Bond, S. (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58(2), 277–297.
- Atkin, D., Khandelwal, A., and Osman, A. (2019): "Measuring Productivity: Lessons from Tailored Surveys and Productivity Benchmarking," *AEA Papers and Proceedings*, 109, 444-449.
- Arnold, J.M., Javorcik, B., Lipscomb, M. and Mattoo, A. (2014): "Services Reform and Manufacturing Performance: Evidence from India," *The Economic Journal*, 126, 1–39.
- Blundell, R.W., and Bond, S. (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87(1), 115–143.
- Cobb, C., and Douglas, P. (1928): "A Theory of Production," *American Economic Review*, 18(1) (Supplement: Papers and Proceedings), 139–165.
- Chen, S., Chernozhukov, V., and Fernández-Val, I. (2019): "Mastering Panel Metrics: Causal Impact of Democracy on Growth", *AEA Papers and Proceedings*, 109, 77-82.
- Gandhi, A., Navarro, S. and Rivers, D. (2020): "On the Identification of Gross Output Production Functions", *Journal of Political Economy*, 128(8), 2973–3016.
- Janes, L., Koelle, M. and Quinn, S. (2022): "Do Capital Grants Improve Microenterprise Productivity?", *Working paper*.

- Levinsohn, J. and Petrin, A. (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 70, 317 – 342.
- McKenzie, D. (2011): "How Can We Learn Whether Firm Policies Are Working in Africa? Challenges (and Solutions?) For Experiments and Structural Models," *Journal of African Economies*, 20(4), 600–625.
- Olley, S. and Pakes, A. (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263 – 1297.
- Petrin, A. and Sivadasan, J. (2013): "Estimating Lost Output from Allocative Inefficiency, with an Application to Chile and Firing Costs," *The Review of Economics and Statistics*, 95(1), 286–301.
- Shenoy, A. (2021): "Estimating the Production Function Under Input Market Frictions," *Review of Economics and Statistics*, 104(4), 666-679.
- Szabo, A. (2016): "Measuring Firm-Level Inefficiencies in the Ghanaian Manufacturing Sector," *Economic Development and Cultural Change*, 66, 447–487.
- Windmeijer, F. (2018): "Testing Over- and Underidentification in Linear Models, with Applications to Dynamic Panel Data and Asset-Pricing Models," *Working paper*.
- Woolridge, J. (2009): "On Estimating Firm-level Production Functions using Proxy Variables to Control for Unobservables," *Economics Letters*, 104(3), 112-114.

2.1 Introduction: From accumulation to production

The previous lecture considered the firm's accumulation problem in an intertemporal context. The theoretical model included a production function, but neither the theory nor the empirical results emphasised the importance of the *production process*. In this lecture, we develop the earlier ideas to consider production; our emphasis will lie primarily upon the use of linear panel data methods to estimate the shape of a production function, and will consider how these methods can help us to understand manufacturing processes in a developing economy.

2.2 Theoretical model: Optimisation under Cobb-Douglas production

> The progressive refinement during the recent years in the measurement of the volume of physical production in manufacturing suggests the possibility of attempting (1) to measure the changes in the amount of labor and capital which have been used to turn out this volume of goods, and (2) to determine what relationships existed between the three factors of labor, capital, and product.
>
> Cobb and Douglas (1928, p.139)

2.2.1 Production without adjustment costs

The Cobb-Douglas production function is a standard part of every economist's toolkit. In this lecture, we will consider the use of Cobb-Douglas production functions to learn about production processes in developing economies. We will focus primarily on empirical methods for estimating such functions with panel data.

We begin, though, by considering a theoretical model of an optimising firm with Cobb-Douglas production; this follows the discussion in sections 2 and 3 of Bond and Söderbom (2005). First, suppose that a firm produces output Y_i, sold at price P, by employing labour L_i at wage W and by hiring capital K_i at rent U. Assume that the firm maximises profits, and that production follows a Cobb-Douglas function with technology A_i:

$$\max_{K_i, L_i} \pi(K_i, L_i; A_i, P, U, W) = P \cdot Y(K_i, L_i; A_i) - U \cdot K_i - W \cdot L_i \tag{2.1}$$

$$Y(K_i, L_i; A_i) = A_i K_i^{\alpha} L_i^{\beta}. \tag{2.2}$$

Of course, this could refer to a single period in a discrete repeated problem. However, note that we begin with the simple case where there are no adjustment costs. In effect, the firm is allowed to optimise in each period *separately*; for this reason, we can specify the firm's problem as though it is a static one.

The first-order conditions are straightforward (where K_i^* and L_i^* denote the optimal choices):

$$U = P \cdot \frac{\partial Y_i(K_i^*, L_i^*; A_i)}{\partial K_i} = P \cdot \alpha A_i K_i^{*\alpha-1} L_i^{*\beta} \tag{2.3}$$

$$W = P \cdot \frac{\partial Y_i(K_i^*, L_i^*; A_i)}{\partial L_i} = P \cdot \beta A_i K_i^{*\alpha} L_i^{*\beta-1}. \tag{2.4}$$

Taking logs (denoted by lower case):

$$u = p + \ln \alpha + a_i + (\alpha - 1) \cdot k_i^* + \beta \cdot l_i^* \tag{2.5}$$

$$w = p + \ln \beta + a_i + \alpha \cdot k_i^* + (\beta - 1) \cdot l_i^*. \tag{2.6}$$

We can, for example, solve for l_i^* by subtracting the bottom line from the top:

$$l_i^* = k_i^* + \ln \beta - \ln \alpha + u - w. \tag{2.7}$$

Substituting into, for example, equation 2.5:

$$u = p + \ln \alpha + a_i + (\alpha - 1) \cdot k_i^* + \beta \cdot k_i^* + \beta \ln \beta - \beta \ln \alpha + \beta u - \beta w. \tag{2.8}$$

Therefore, as in Bond and Söderbom:

$$\begin{aligned} k_i^* = &\left(\frac{1-\beta}{1-\alpha-\beta}\right)\cdot \ln\alpha + \left(\frac{\beta}{1-\alpha-\beta}\right)\cdot \ln\beta \\ &- \left(\frac{1-\beta}{1-\alpha-\beta}\right)\cdot (u-p) - \left(\frac{\beta}{1-\alpha-\beta}\right)\cdot (w-p) + \left(\frac{1}{1-\alpha-\beta}\right)\cdot a_i; \end{aligned} \tag{2.9}$$

$$\begin{aligned} l_i^* = &\left(\frac{\alpha}{1-\alpha-\beta}\right)\cdot \ln\alpha + \left(\frac{1-\alpha}{1-\alpha-\beta}\right)\cdot \ln\beta \\ &- \left(\frac{\alpha}{1-\alpha-\beta}\right)\cdot (u-p) - \left(\frac{1-\alpha}{1-\alpha-\beta}\right)\cdot (w-p) + \left(\frac{1}{1-\alpha-\beta}\right)\cdot a_i. \end{aligned} \tag{2.10}$$

Three points immediately deserve noting:

(i) *Both* k_i^* and l_i^* are functions of the firm-specific technology shock, a_i. This is the basis for 'transmission bias', which we consider shortly.

(ii) In the special case that $\alpha + \beta = 1$ — that is, under constant returns to scale — the firm size (that is, the optimal choice of k_i^* and l_i^*) is indeterminate. That is, the firm could solve its optimisation problem solely in 'intensive form' (that is, expressed 'per worker').

(iii) The production function is additively separable in logs. This assumption is not necessarily unreasonable, of course — but it is an implicit restriction on the production process (which can be relaxed, for example, by using a 'translog' production function, as we will discuss shortly).

But there is a more fundamental point still. Consider the original production function, expressed in logs at the optimal capital and labour:

$$y_i(k_i^*, l_i^*; a_i) = \alpha \cdot k_i^* + \beta \cdot l_i^* + a_i. \tag{2.11}$$

Therefore, as Bond and Söderbom explain (page 1):

> If all inputs are chosen optimally and are perfectly flexible in the sense that they can be varied immediately without incurring any costs, then all inputs are perfectly collinear with the productivity shocks observed by firms... the parameters on the perfectly flexible inputs are not identified, regardless of the estimation technique considered.

That is, even if we had a *huge* number of observations, measured without error, we *still* couldn't learn the parameters α and β — "*regardless of the estimation technique considered*". This clearly presents a conceptual challenge: is there *any* hope for understanding the structure of firm production?

2.2.2 Introducing adjustment costs

Bond and Söderbom proceed to consider adjustment costs. They specify a value function — as in Lecture 1, we see the *value* of a value function in thinking about a firm's intertemporal problem — that allows for costly adjustment:

$$V_t(K_{t-1}, L_{t-1}) = \max_{I_t, H_t} P_t F_t(K_t, L_t) - P_t^K I_t - P_t^K G_t(I_t, K_t) - W_t L_t - W_t C_t(H_t, L_t) + \psi_t \cdot \mathbb{E}(V_{t+1}(K_t, L_t)) \quad (2.12)$$

$$\text{subject to } K_t = (1-\delta)K_{t-1} + I_t \quad (2.13)$$

$$L_t = (1-q)L_{t-1} + H_t, \quad (2.14)$$

where I_t refers to *gross investment*, H_t is *gross hiring*, δ is the *depreciation rate* and q is the labour *quit rate*. Critically, note the introduction of functions $G_t(I_t, K_t)$ and $C_t(H_t, L_t)$; these represent *adjustment costs* for capital and labour respectively.

The authors show how, under this structure, the chosen values of capital and labour are no longer collinear with productivity shocks; in effect, the parameters α and β may be identified if the endogeneity problem can be resolved (as we will discuss shortly). I leave you to consider their discussion of first-order conditions; in these notes, I emphasise simply the authors' illustrative example (page 10):

> ... consider the firm's response to a large, permanent increase in productivity, assuming that adjustment costs for capital are higher than those for labour. Eventually the firm will want to have higher levels of both inputs, as in the case where inputs are flexible. However the capital stock will adjust more slowly than employment since adjusting the capital stock is relatively expensive. This will generate a lower capital-labour ratio during the period after the shock when significant adjustments are occuring... the combination of productivity shocks with different levels of adjustment costs for different inputs will generate variation both across firms and over time in, for example, the capital-labour ratio.

2.3 Empirical method: Linear dynamic panel estimation

2.3.1 The problem of time-varying unobservables

So we know that, so long as there are adjustment costs, we *can* identify the parameters α and β — but *how*? That is the question that we now consider. As before, we assume a Cobb-Douglas production function (expressed in logs); I shift notation to follow Eberhardt and Helmers:

$$y_{jt} = \beta_0 + \beta_k \cdot k_{jt} + \beta_l \cdot l_{jt} + \varsigma_{jt}. \quad (2.15)$$

We can decompose the error term as:

$$\varsigma_{jt} = \eta_j + \omega_{jt} + \gamma_t + v_{jt}, \quad (2.16)$$

where η_j represents firm j's productivity deviation from some reference firm, γ_t represents average productivity changes over time, ω_{jt} represents a firm- and time-specific productivity shock and v_{jt} captures measurement error. We can therefore rewrite equation 2.15:

$$y_{jt} = \beta_0 + \beta_k \cdot k_{jt} + \beta_l \cdot l_{jt} + \eta_j + \omega_{jt} + \gamma_t + v_{jt}. \tag{2.17}$$

Equation 2.17 illustrates nicely the fundamental identification problem in this context. As Eberhardt and Helmers explain it (page 6, emphasis in original):

> The main problem for estimation of specifications such as [Equation 2.17] arises from the plausible suggestion that firms decide on their choice of inputs (l, k) based on the realized firm-specific productivity shock (ω_{jt}), *which only they observe*: for instance a favorable productivity shock to firm j might induce higher levels of investment... Since ω_{jt} is suggested to 'transmit to' the input choices, this particular problem is known as the 'transmission bias'.

This is precisely the point noted earlier: that k^* and l^* are likely to be a function of the (unobservable) firm-specific technology shock. It is also a very familiar endogeneity problem; it is directly analogous, for example, to the problem that individuals with higher 'unobserved ability' may be likely to attend school for longer. You will be familiar with a number of methods for dealing with endogeneity in the schooling context; for example, a randomised controlled trial, instrumental variable methods, a regression discontinuity design, *etc*. All of these methods could, in principle, be available in this context — so, for example, a researcher might be able to use factor prices as instrumental variables for endogenous labour and capital.

However, there are at least two important practical distinctions between identification of a firm production function and identification of a Mincerian earnings regression:

(i) It may be practically difficult to find any plausible exogenous variation — or, indeed, use an RCT to generate such variation — particularly in the case of larger firms.[5] This creates important *practical problems* for using traditional experimental or quasi-experimental methods.

(ii) Unlike an individual's choice of education, a firm's choice of capital and labour is likely to *vary substantially over time*. This creates an important *practical opportunity* to use such variation for identification.

It would be tempting to suggest relying upon a fixed-effect estimator (sometimes known

[5] For a really interesting discussion of the problems of randomising in order to learn about firms in Sub-Saharan Africa — and, in particular, the implications of heterogeneity — see McKenzie (2011).

as the 'within-group estimator'). However, remember that a fixed-effect estimator is appropriate where the endogenous variables are correlated to *time-invariant* unobservable — the *fixed* effect. Our problem, however, is more complicated: remember that our endogenous variables (l and k) are correlated to the firm-specific productivity shock (ω_{jt}), and this is a *time-varying unobservable*. For this reason, we *cannot* rely upon a fixed-effect estimator to identify β_k and β_l.

We *can*, however, use a class of estimators known as 'linear dynamic panel estimators'. These estimators are sometimes referred to as relying upon 'own-instrumentation', for reasons that will soon be obvious. We follow Eberhardt and Helmers by ignoring measurement error ($v_{jt} = 0$), and by making a specific assumption about the autoregressive form of ω_{jt}. There is substantial flexibility in the form of this assumption, but we will take a simple case for its illustrative value.

Assumption 5 (The autoregressive structure of technology shocks)
Firm-specific technology shocks follow a first-order autoregressive process:

$$\omega_{jt} = \rho \cdot \omega_{j,t-1} + \xi_{jt}, \tag{2.18}$$

where $|\rho| < 1$ *and* $\xi_{jt} \sim MA(0)$.

By rearranging, this implies:

$$\omega_{jt} = \rho \cdot (y_{j,t-1} - \beta_0 - \beta_k \cdot k_{j,t-1} - \beta_l \cdot l_{j,t-1} - \eta_j - \gamma_{t-1}) + \xi_{j,t} \tag{2.19}$$

$$\Leftrightarrow y_{jt} = \rho y_{j,t-1} + \beta_l l_{jt} - \rho\beta_l \cdot l_{j,t-1} + \beta_k k_{jt} - \rho\beta_k k_{j,t-1} + (1-\rho) \cdot (\beta_0 + \eta_j) + (\gamma_t - \rho\gamma_{t-1}) + \xi_{jt}. \tag{2.20}$$

Following Eberhardt and Helmers, we can rewrite this as:

$$y_{jt} = \pi_1 y_{j,t-1} + \pi_2 l_{jt} + \pi_3 l_{j,t-1} + \pi_4 k_{jt} + \pi_5 k_{j,t-1} + \alpha_j^\star + \gamma_t^\star + \xi_{jt}, \tag{2.21}$$

$$\text{where } \alpha_j^\star \equiv (1-\rho)(\beta_0 + \eta_j) \tag{2.22}$$

$$\gamma_t^\star \equiv \gamma_t - \rho\gamma_{t-1} \tag{2.23}$$

$$\text{and } \pi_3 = -\pi_1\pi_2 \tag{2.24}$$

$$\pi_5 = -\pi_1\pi_4. \tag{2.25}$$

Note that equations 2.22 and 2.23 define *identities*; they just allow us to write equation 2.21 more simply. But equations 2.24 and 2.25 are substantive *restrictions* upon our model: we can (i) estimate equation 2.21 *without* imposing these restrictions, then (ii) estimate *with* the restrictions, and then (iii) test whether the restrictions are valid. In effect, equations 2.24 and 2.25 impose a stability on the production process across time periods; for this reason, they are sometimes referred to as the *common factor restrictions*.

We're not done yet. Were we to estimate equation 2.21 directly — that is, by OLS — we would face a basic endogeneity problem: our error term would include $\alpha_j^\star$, a firm fixed

effect, which is likely to correlated with l_{jt}, $l_{j,t-1}$, k_{jt} and $k_{j,t-1}$. We can, however, use first differences (where, for any x, I define $\Delta x_t \equiv x_t - x_{t-1}$):

$$\Delta y_{jt} = \pi_1 \Delta y_{j,t-1} + \pi_2 \Delta l_{jt} + \pi_3 \Delta l_{j,t-1} + \pi_4 \Delta k_{jt} + \pi_5 \Delta k_{j,t-1} + \Delta \gamma_t^\star + \Delta e_{jt}. \tag{2.26}$$

But we're *still* not done! Remember that e_{jt} includes the firm-specific productivity shock (ξ_{jt}), which we anticipate to be correlated with k_{jt} and l_{jt}; it follows that Δe_{jt} is correlated with Δl_{jt}, $\Delta l_{j,t-1}$, Δk_{jt} and $\Delta k_{j,t-1}$. (Additionally, you should be able to *show* directly that Δe_{jt} correlates with $\Delta y_{j,t-1}$; this should follow immediately from equation 2.21.) This, in short, is precisely the problem of *time-varying* unobservables.

Here, at last, enters the fundamental idea of linear panel data estimators: having taken the first difference, we can use *lagged values of the endogenous variables* as valid instruments. Let's start with y_{jt} — and, for simplicity, we will ignore measurement error: $v_{jt} = 0 \; \forall \; t$.

2.3.2 Lagged levels as instruments for differences

Note that we can write $y_{jt} = C_1 + e_{jt}$; it follows that

$$\mathbb{E}(y_{j,t-1} \cdot \Delta e_{jt}) = \mathbb{E}[y_{j,t-1} \cdot (e_{j,t} - e_{j,t-1})] = -\mathbb{E}(y_{j,t-1} \cdot e_{j,t-1}) \neq 0. \tag{2.27}$$

That is, $y_{j,t-1}$ is *not* a valid instrument. However, what about $y_{j,t-2}$?

$$\mathbb{E}(y_{j,t-2} \cdot \Delta e_{jt}) = \mathbb{E}[y_{j,t-2} \cdot (e_{j,t} - e_{j,t-1})] = 0. \tag{2.28}$$

Therefore, because of our assumption on the error process in equation 2.18, it follows that $y_{j,t-2}$ *is* a valid instrument — and, by the same logic, so too are further lags. We can refer to Equation 2.28 as a 'moment condition' — and, by using further lags, we can write more moment conditions for lagged output:

$$\mathbb{E}[y_{j,t-s} \cdot \Delta e_{jt}] = 0 \qquad \text{for } t = 3, \ldots, T \text{ and } s \geq 2. \tag{2.29}$$

Similarly, we can use lags of k_{jt} and l_{jt} as instruments too. But exactly *which* lags we can use depends critically upon our assumptions about how labour and capital are chosen.

Assumption 6 (Timing of labour choices) *For any given time period, labour is chosen* after *the firm has observed that period's shock. That is,*

$$\mathbb{E}(l_{j,t+1} \cdot e_{jt}) \neq 0 \tag{2.30}$$

$$\mathbb{E}(l_{jt} \cdot e_{jt}) \neq 0 \tag{2.31}$$

$$\mathbb{E}(l_{j,t-1} \cdot e_{jt}) = 0. \tag{2.32}$$

We can describe this assumption as implying that labour is '***endogenous***'. (Note that this is a more specific sense of the term 'endogenous' than we usually use.) This implies the following moment conditions:

$$\mathbb{E}[l_{j,t-s} \cdot \Delta e_{jt}] = 0 \qquad \text{for } t = 3, \ldots, T \text{ and } s \geq 2. \tag{2.33}$$

For capital, however, we will assume a longer implementation lag.

Assumption 7 (Timing of capital choices) *For any given time period, capital is chosen* before *the firm has observed that period's shock, but* after *observing the shock from the previous period. That is,*

$$\mathbb{E}(k_{j,t+1} \cdot e_{jt}) \neq 0 \tag{2.34}$$
$$\mathbb{E}(k_{jt} \cdot e_{jt}) = 0 \tag{2.35}$$
$$\mathbb{E}(k_{j,t-1} \cdot e_{jt}) = 0. \tag{2.36}$$

We can describe this assumption as implying that capital is '***predetermined***'. As you will see, the basic asymmetry between the capital and labour assumptions is the difference between equations 2.31 and 2.35. This implies one *more* moment condition than we have for labour or output:

$$\mathbb{E}[k_{j,t-s} \cdot \Delta e_{jt}] = 0 \qquad \text{for } t = 2, \ldots, T \text{ and } s \geq 1. \tag{2.37}$$

Typically, we therefore have a large number of valid instruments. For example, suppose that we have a four-period panel ($T = 4$). It follows that we can use the following variables as instruments:

$$\begin{array}{ll} \textit{for } t = 2\text{:} & k_{j1}; \\ \textit{for } t = 3\text{:} & k_{j1}, y_{j1}, l_{j1}, k_{j2}; \\ \textit{for } t = 4\text{:} & k_{j1}, y_{j1}, l_{j1}, k_{j2}, y_{j2}, l_{j2}, k_{j3}. \end{array}$$

We can therefore estimate β_l and β_k consistently by estimating equation 2.26 using as instruments appropriate lags of y_{jt}, k_{jt} and l_{jt}. We will leave aside discussion of exactly *how* we could do this; in short, we could use Two-stage Least Squares ('2SLS'), but would probably use a more sophisticated Generalised Method of Moments ('GMM').

You will recall that a variable needs to satisfy *two* conditions in order to be an instrumental variable: (i) it needs to be *valid*, *and* (ii) it needs to be *informative*. We have discussed only *validity* — and, indeed, this has been the traditional concern of dynamic panel models. The issue of informativeness in dynamic panel models is likely to become more important in the near future, following recent work by Windmeijer (2018).

2.3.3 Lagged differences as instruments for levels

The previous section explained how we can use *lagged levels* as instruments for endogenous *differences*. This is sometimes referred to as an 'Arellano-Bond estimator', following the seminal work of Arellano and Bond (1991); additionally, we may refer to it as the 'Difference GMM' approach.

In some circumstances, however, we can do better — we can also use *lagged differences* as instruments for endogenous *levels*. This creates a *system* of two estimating equations, and this estimation method is therefore often referred to as 'System GMM' (alternatively, it is sometimes referred to as a 'Blundell-Bond estimator', following Blundell and Bond (1998).) This is discussed in more detail by Eberhardt and Helmers.

2.3.4 Specification tests

This own-instrumentation strategy has required several strong assumptions. It is important, therefore, to consider how we might *test* these restrictions. We will consider three types of test.

(i) ***The common factor restriction:*** We noted earlier that the common factor restrictions (in equations 2.24 and 2.25) can be tested.

(ii) ***The autoregressive structure of*** ω_{jt}***:*** Equation 2.18 assumes an AR(1) structure for ω_{jt}. In turn, this implies:

  (a) Δe_{jt} *should* be serially correlated between any t and $t-1$;

  (b) Δe_{jt} *should not* be serially correlated at any order greater than one (for example, between any t and $t-2$).

  We can test both of these implications. Note that the *first* implication should lead us to *reject* a null hypothesis of 'no first-order serial correlation', but that the *second* implication should lead us *not to reject* any null hypothesis of no serial correlation at higher orders.

(iii) ***The Sargan-Hansen overidentification test:*** We earlier considered a hypothetical case in which we have a four-period panel; we showed that, across periods $t = 2$, $t = 3$ and $t = 4$, this implies 12 instruments. We would use these instruments to estimate equation 2.26; this equation has *five* endogenous variables (*i.e.* $\Delta y_{j,t-1}$, Δl_{jt}, $\Delta l_{j,t-1}$, Δk_{jt} and $\Delta k_{j,t-1}$). The number of instrumental variables is larger than the number of endogenous variables: we can therefore describe our parameters as *over-identified*.

  Over-identification generally implies that we can test the validity of our instruments. In effect, we are asking, '*Assuming that* ***some*** *of our instruments are valid, are*

***all** of them valid?'*. In practice, this involves finding a special kind of weighted average of the *sample moment conditions* — that is, the sample averages corresponding to the moment conditions in equations 2.29, 2.33, 2.37. We refer to this statistic as the 'Sargan-Hansen statistic'; it is discussed in more detail by Eberhardt and Helmers.

If all of the instruments are valid, the value of the Sargan/Hansen statistic should be *small*; that is, the sample averages should all be close to zero. But if *any* of the instruments are invalid, the Sargan/Hansen statistic should be relatively *large*. Specifically, if all of the instruments are valid, we can say that $S \sim \chi^2(r-k)$, where S refers to the Sargan-Hansen statistic, r is the number of instrumental variables and k is the number of endogenous variables.

Consider Table 2.1; this is an excerpt from a table in a 2010 NBER Working Paper about determinants of health outcomes. (I have extracted the two columns reporting GMM estimates from a linear panel data model, and I have extracted only the diagnostic statistics.) Look at the reported Sargan/Hansen statistics: 79.62 (with 54 degrees of freedom) and 132.09 (with 75 degrees of freedom). The statistics respectively imply $p \approx 0.0132$ and $p < 0.0000$. If the data could talk, it would scream, "your identifying assumptions are *wrong*; these instrumental variables are *not* all valid!". That is, the Sargan/Hansen statistics *should* have prompted the authors to try a different set of instruments. And this is the final important point that we will emphasise about linear panel estimation: that we can use different sets of *instruments* to implement different sets of *identifying assumptions* — assumptions that can then be tested. For example, for illustrative purposes, we earlier assumed that labour is endogenous, that capital is predetermined and that ω_{jt} follows an AR(1) process; however, these are falsifiable assumptions that can — and should — be tested.

Table 2.1: **An excerpt from a table in an NBER Working Paper**

Instruments	Differenced Eq: Lagged Z_i and Δ community variables Levels Eq: Lagged difference for Z_i for (t-1) and prior	
Number of Instruments used	75	99
Overall Statistic	$\chi^2(20)$= 247.93**	$\chi^2(23)$= 376.41**
Sargan/Hansen test of over-identification	$\chi^2(54)$= 79.62**	$\chi^2(75)$= 132.09**
Test for Autocorrelation		
AR(1) in first differences (*z*-statistic)	-3.595 **	-3.934 **
AR(2) in first differences (*z*-statistic)	-0.838	-0.422
N	8,645	8,642
Unique individuals	4,120	4,120

2.4 A different method...

In this lecture, we are focusing on dynamic panel methods for estimating the parameters of production functions. However, there is an alternative method that is also very popular in applied work: namely, the use of proxy variables to control for endogeneity of input choice. Olley and Pakes (1996) use a flexible function of firm investment to proxy for the firm technology shock. Levinsohn and Petrin (2003) instead use a flexible function of intermediate inputs. Wooldridge (2009) shows how these methods can be estimated using GMM. Most recently, Ackerberg, Caves and Frazer (2015) use investment and/or intermediate inputs conditional upon labour demand.

You are not required to understand the details of these approaches for this course. However, you should read the excellent overview of this literature summarised by Ackerberg *et al*. You should particularly note Ackerberg *et al*'s comparisons between the proxy variable method and the dynamic panel approach (pages 16 and 17):

> The DP approach does not need the assumptions that generate invertibility of the variable input demand function. So, e.g., it can allow for unobserved cost shocks to *all* inputs, unlike our approach, which does not allow such shocks to the price of [intermediate inputs]. On the other hand, the DP derivation seems to rely on the linearity of the ω_{it} process — in contrast, OP, LP, and our approach can treat the first-order markov process completely nonparametrically. There are other differences between the models. For example, the DP literature can be extended to allow for a fixed effect α_i in addition to the AR(1) process, while generally speaking, this is challenging in our context because it would tend to violate the scalar unobservable assumption. The dynamic panel literature can also potentially allow future values of the intermediate input or investment variable to depend on past ϵ_{it}'s, while our approach cannot. On the other hand, as elaborated on in OP, the scalar unobservable assumption of OP/LP and our approach makes it failry straightforward to extend the methodologies to address endogenous exit (selection) from a sample — this would be considerably harder in the DP context. In summary, both approaches require strong (but different) assumptions. In some cases, a-priori beliefs about a particular production process and/or data considerations may guide choices between the two approaches. In other cases, one may want to try both techniques. Finding that estimates are consistent across multiple techniques with different assumptions is surely more convincing than using only one.

2.5 The role of intermediate inputs

It is worth pausing for a moment to recognise the breadth of the underidentification critique that we considered earlier. In this lecture, we consider just two types of inputs: labour and capital. As the previous section set out, we will assume adjustment costs for

both of those factors.

But what about factors for which there is *no* adjustment cost? These factors — such as raw materials, energy and services — are typically referred to as 'intermediate inputs'. A recent paper by Gandhi, Navarro and Rivers (2020) emphasises that the presence of completely flexible inputs causes fundamental problems for the identification of production functions.

To be sure, Gandhi *et al* agree that, as we just discussed, the use of dynamic panel methods is appropriate *if* there are no flexible inputs. However, as the authors argue (page 21),

> ... the bulk of empirical work based on production function estimation has focused on environments in which some inputs are quasi-fixed (namely capital and labor) *and* some inputs are flexible. It is this setting that motivates our problem and distinguishes our approach from the dynamic panel literature.

Gandhi *et al* propose an alternative non-parametric method, based in part on an estimation of the firm's first-order condition. Their method is intricate, and lies beyond the scope of this course; nonetheless, it is imperative that you read their critique if you are interested in estimating production functions in your own empirical work.[6] More recently, Shenoy (2020) provides a selection criterion for choosing between the Gandhi *et al* estimator and the autoregressive approach that we have considered in this lecture.

2.6 Quantities or revenues?

Up to this point, we have discussed inputs and we have discussed outputs – without ever really being precise about what we are measuring, nor how. In particular, we have not discussed whether these inputs and outputs are measured as *quantities* or as *expenditures and revenues*. Given our particular focus, this is not a topic that we will spend time on in the lecture. Nonetheless, it is an important and interesting issue, and one that you should be aware of if estimating production functions in your own work.

Atkin et al (2019) explain very nicely the distinction between the two key concepts, and the practical implication:

> What the researcher typically wants is a measure of physical output conditional on physical inputs, termed quantity-based productivity (TFPQ). This requires data on input and output quantities that are not typically available. In cases where these data are available, quantities are likely measured with substantial error since they cannot be easily read off accounting statements.

[6] If you interested in estimating production functions in your own empirical work, you may also wish to look at the recent work of Chen *et al* (2019), who propose debiasing procedures for linear panel estimators.

Even if well measured, product specifications and quality levels can vary dramatically across firms and within firms across product lines—variation that is not well captured by disaggregated product categories in typical administrative datasets. This makes it difficult to both measure productivity for firms that produce many varieties or to compare productivity across firms making different varieties. Multi-product firms pose further challenges since output and input mixes vary even more widely across products than within.

As most firm-level datasets only provide expenditure and revenue data, much of the literature relies on revenue-based productivity (TFPR) measures that also capture differences in markups and quality across firms. However, if a firm's capabilities come from its ability to produce both quality and quantity, TFPR may be closer to the object of interest even though it confounds forces unrelated to productivity.

2.7 Empirical results: Firm size analysis using production functions

Söderbom and Teal (2004) concerns the role of firm size in determining firm production practices, using a six-year panel dataset of Ghanaian manufacturing firms. (In total, the authors use a total of 676 observations on 143 firms.) In order to do this, the authors were required to estimate a firm production function. We will return in Lecture 4 to consider Söderbom and Teal's analysis of firm size; in this lecture, we will discuss their analysis of the production function.

Söderbom and Teal's basic specification is a *translog* production function; using our earlier notation, we can write this as:

$$y_{it} = \beta_0 + \beta_k \cdot k_{it} + \beta_l \cdot l_{it} + \frac{1}{2} \cdot \delta_{kl} k_{it} l_{it} + \frac{1}{2} \cdot \delta_{kk} k_{it}^2 + \frac{1}{2} \cdot \delta_{ll} l_{it}^2 + \eta_i + \gamma_t + \omega_{it}. \tag{2.38}$$

Note that, under this translog specification,

$$\frac{\partial y_{it}(k_{it}, l_{it})}{\partial k_{it}} = \beta_k + \frac{1}{2} \cdot \delta_{kl} l_{it} + \delta_{kk} k_{it} \tag{2.39}$$

$$\frac{\partial y_{it}(k_{it}, l_{it})}{\partial l_{it}} = \beta_l + \frac{1}{2} \cdot \delta_{kl} k_{it} + \delta_{ll} l_{it}. \tag{2.40}$$

That is, the translog function allows the marginal product of capital to *vary* with the labour input — and *vice versa* the marginal product of labour. In this way — unlike the Cobb-Douglas function — it allows for estimation *without* imposing separability. However, note that it *nests* the Cobb-Douglas as a special case where $\delta_{kl} = \delta_{kk} = \delta_{ll} = 0$; this, of course, is a testable restriction. In their translog estimations, Söderbom and Teal do not reject $H_0 : \delta_{kl} = \delta_{ll} = \delta_{kk} = 0$, so they rely upon the Cobb-Douglas estimates. So shall we; for illustrative purposes, we will consider only the authors' estimation of

the Cobb-Douglas function. Table 2.2 collates the authors' estimation results for Cobb-Douglas production of value-added. Note that the GMM specification uses "lags of the explanatory variables in levels as instruments for contemporaneous differences, and lags of the explanatory variables expressed in first differences as instruments for contemporaneous levels" (p.379).[7]

In this lecture, we will consider Söderbom and Teal's results by a class discussion of the following questions.

(i) Table 2.2 reports three estimations. Which should we prefer? Why?

(ii) Suppose that an urban Ghanaian manufacturing firm were to increase its labour force by 10%. *Ceteris paribus*, how much should we expect its output to increase? That is, what would be the anticipated *causal effect* of such a change?

(iii) The 'other firm controls' are described in the notes to Table 2 in the original paper: they are controls 'for the age of the firm, industry, ownership structure and location'. Why are these controls not included in the FE or GMM estimations?

(iv) Write the hypothesis tested by the 'constant returns to scale' diagnostic.

(v) The GMM estimation produces a tiny p-value for one of the diagnostic tests. Is this a reason for a researcher to reject the GMM estimates?

(vi) Which of the diagnostic tests might be described as a 'test of overidentifying restrictions'?

(vii) There is another diagnostic test for estimating firm production functions using panel data, which the authors do not report. What is it?

2.8 Production function estimation as a foundation for other analysis

In this lecture, we have discussed the estimation of firm production functions almost as though the exercise is an end in itself. Of course, the parameters of production functions *are* inherently interesting — for example, given the implications for understanding how costs are likely to be divided between labour and capital, or for thinking about the likely effect of, say, changes in wages or rental costs. However, in most cases, we estimate firm production functions as a first step towards some further empirical question. In the case of Söderbom and Teal (2004), for example, the ultimate research question concerns firm size and efficiency; we will return to discuss their results on this issue in Lecture 4. Similarly, Petrin and Sivadasan (2013) use the proxy variable method (summarised previously) to estimate the extent of allocative inefficiencies in Chilean manufacturing firms; Szabo (2016) conducts a similar exercise to measure potential gains from adjusting input use among Ghanaian manufacturing firms. Conversely, in the large literature on

[7] That is, the 'GMM' specification uses 'System GMM'.

Table 2.2: **Results from Söderbom and Teal (2004), Tables 2 and 3**

	Dependent variable: Log value-added		
	OLS	**FE**	**GMM**
	(1)	(2)	(3)
Log employment	**0.89****	**0.34**	**0.73****
Log capital	**0.18****	**-0.22**	**0.31****
Human capital controls	✓	✓	✓
Time dummies	✓	✓	✓
Other firm controls	✓	✗	✗
R^2	0.74	0.09	
Constant returns to scale (p-value)	0.39	0.04**	0.85
Residuals uncorrelated: $t, t-1$ (p-value)			0.00***
Residuals uncorrelated: $t, t-2$ (p-value)			0.93
Sargan-Hansen (p-value)			0.79

Confidence: *** $\leftrightarrow$ **99%**, ** $\leftrightarrow$ **95%**, * $\leftrightarrow$ **90%**.

productivity, production function estimation is often used to recover predicted values of TFP — which can then be regressed on other explanatory variables. For example, this is the approach used by Arnold, Javorcik, Lipscomb and Mattoo (2014) in analysing the role of policy reforms in India; the authors first implement the method of Ackerberg *et al*, then regress predicted TFP on various measures of liberalisation. Similarly, Janes *et al* (2020) take a similar approach to analyse the effect of capital drop experiments (as discussed in our previous lecture) on microenterprise productivity.

3 Lecture 3: The Firm and Technology Adoption

References:

- ⋆ Griliches, Z. (1957): "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 25(4), 501–522.
- ⋆ Roy, A.D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–146.
- ⋆ Michler, Tjernström, Verkaart, and Mausch (2018): "Money Matters: The Role of Yields and Profits in Agricultural Technology Adoption," *American Journal of Agricultural Economics*, https://doi.org/10.1093/ajae/aay050.
- ⋆ Suri, T. (2011): "Selection and Comparative Advantage in Technology Adoption," *Econometrica*, 79(1), 159–209.
- Atkin, D., Khandelwal, A., and Osman, A. (2017): "Exporting and Firm Performance: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 132(2), 551–615.
- Card, D. (1996): "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64(4), 957–979.
- Duflo, E., Kremer, M., and Robinson, J. (2011): "Nudging Farmers to use Fertilizer: Theory and Experimental Evidence from Kenya," *American Economic Review*, 101(6), 2350–2390.
- Foster, A.D. and Rosenzweig, M.R. (1995): "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy*, 103(6), 1176–1209.
- Hanna, R., Mullainathan, S., and Schwartzstein, J. (2014): "Learning Through Noticing: Theory and Experimental Evidence in Farming", *Quarterly Journal of Economics*, 1311-1353.
- Lemieux, T. (1998): "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection," *Journal of Labor Economics*, 16(2), 261–291.

3.1 Introduction: 'Origins, slopes, and ceilings'

The first two lectures have taken firm production technology as given. This may be reasonable if we are treating technology as fixed — as in the first lecture — or if we see it as evolving according to productivity shocks — as in the second lecture. But, of course, the processes of firm technology development can be much more purposive: innovation can be driven by the costs and benefits of innovating. That is what we shall consider in this lecture.

3.1.1 Griliches (1957): "Hybrid corn: An exploration in the economics of technological change"

> ... the process of innovation, the process of adapting and distributing a particular invention to different markets and its acceptance by entrepreneurs, is amenable to economic analysis.

So argued Griliches (1957, p.522) in what was, apparently, a radical claim for its time.[8] Griliches' study — of the adoption of hybrid maize seed across the United States — is captured dramatically by his Figure 1 (reproduced here as Figure 3.1). The graph shows (i) that different regions adopted hybrid maize at different times and rates, and (ii) that the path of adoption followed an 'S' shape, conveniently captured by the logistic function.

Figure 3.1: **Figure 1 from Griliches (1957)**

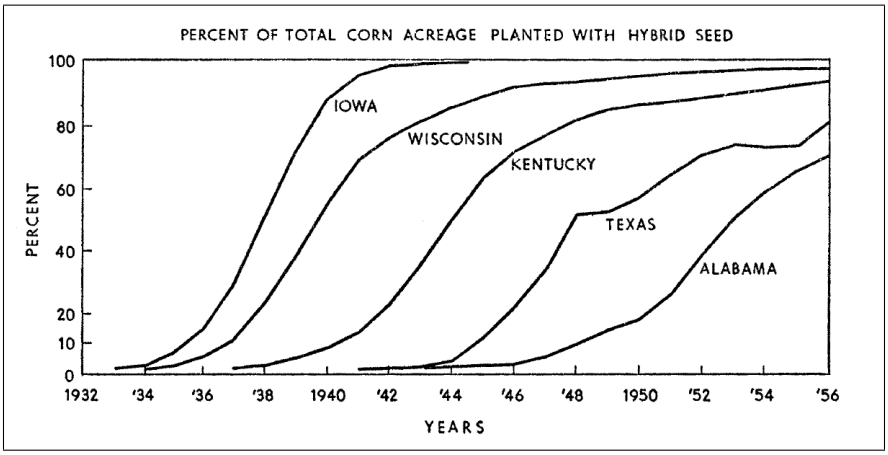
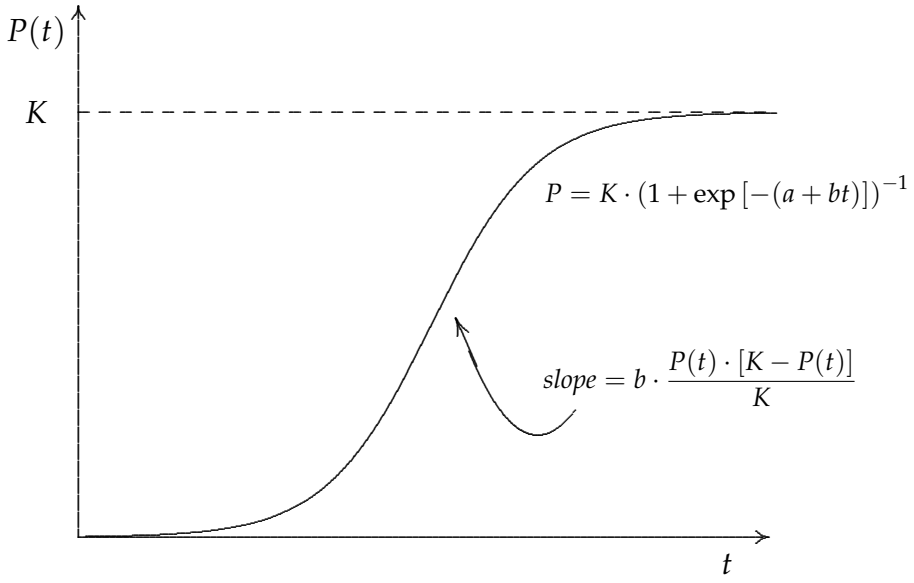


Figure 3.2 shows a logistic function; the accompanying formula shows that the curve can be completely described by (i) a location parameter (a), (ii) a rate parameter (b) and (iii) an asymptotic maximum value (K) — or, as Griliches put it, 'origins, slopes, and ceilings'. In short, Griliches argued — in sections 3, 4 and 5 of his paper respectively — that variations in origins, slopes and ceilings can all be explained significantly by variation in the expected pay-off of the technology adoption.

In some respects, this may seem a trivial observation: after all, why *shouldn't* a firm weigh economic costs and benefits when deciding whether or not to adopt a new technology? However, Griliches' results show that expected returns can drive not only firms' choice of *factors of production*, but also of their choice of *technology of production*. Further, Griliches' work emphasises the importance for technology adoption of *heterogeneity* across different

[8] See, for example, Griliches' cutting final footnote!

Figure 3.2: **A logistic curve**



firms. Griliches' Figure 1 (our Figure 3.1), for example, begs at least two fundamental questions. First, why did some *states* adopt hybrid corn differently to other states? This, essentially, is answered by Griliches' analysis of 'origins, slopes and ceilings'. But there is a second question, too: even within each state, why did some *individual farmers* adopt corn earlier than others?

This second question has prompted a large literature, reaching far beyond economics — and, indeed, reaching back before Griliches' work. For example, this is the central question in the 1962 book *Diffusion of Innovations*, which apparently introduced the term 'early adopter'. The book, which cites Griliches' 1957 paper, was written by the sociologist Everett Rogers — who grew up on a farm in Iowa, where he saw his father's non-hybrid maize wilting next to the neighbour's hybrid breed.[9]

3.1.2 Roy (1951): "Some thoughts on the distribution of earnings"

Roy's 1951 paper is another classic early work, and another classic early work emphasising the importance of heterogeneity — this time, the importance of heterogeneity for choosing an occupation in the labour market. Roy sketches several scenarios concerning the choice between being a hunter and a fisherman, where the latter represents a 'skilled' occupation (page 137):

> The rabbits are plentiful and stupid and even the less skilled man can ensnare a fair number in a year's hunting while the exercise of a quite appreciable degree of skill does not enable the better hunters to catch many more. The trout, on the other hand, are particularly wily and fight hard, so that many

[9] See Singhal, A. (2005): "Brief Biography of Everett M. Rogers," *Journal of Health Communication*, 10(4), 287–288, p.287.

> men would undoubtedly starve if they had to eat only what they themselves caught; but nevertheless the real fisherman can obtain very big catches in a year's fishing, although such catches are pretty rare occurrences.

In this way, Roy emphasised that different workers have different productivities — and that this heterogeneity determines workers' choice of occupation. But Roy went further; he also emphasised that workers need not have the *same* productivity across hunting and fishing, and that the population distribution of these separate skills need not be independent (again, page 137):

> There may be a marked positive association so that the best hunters and fishermen are, generally speaking, the same people, or the association may be negative so that the best hunters generally make the worst fishermen and conversely.

I leave it to you to read the rest of Roy's parable and its implications. More than 60 years on, it is these basic themes — of unobserved heterogeneity in productivity, and of productivity driving switching decisions — that resonate in Tavneet Suri's recent analysis of hybrid seed adoption by Kenyan farmers. We will spend the rest of this lecture discussing that paper.

3.2 Model: Technology adoption under heterogeneous returns

Suri is concerned to investigate an 'empirical puzzle': if *average* returns to farm technology adoption are so high, why doesn't every farmer adopt? Like Griliches and Roy before her, Suri emphasises the role of heterogeneous returns: she argues that "farmers with high net returns to the technology adopt it and the farmers with low returns do not... adoption decisions are on the whole rational" (p.161).

3.2.1 The basic idea: Endogenous switching into new technology

Suri begins with a theoretical model of seed adoption. Profits under *hybrid seed* equal production revenue less the cost of obtaining such seed:

$$\pi_{it}^{H} = p_{it} Y_{it}^{H} - (b_t s_{it} + a_{it}), \tag{3.1}$$

and profits under *non-hybrid seed* equal production revenue less the cost of replanting non-hybrid seed from the previous year's harvest (which, for simplicity, we treat as zero):

$$\pi_{it}^{N} = p_{it} Y_{it}^{N}. \tag{3.2}$$

Note that, for simplicity, we are excluding here any other inputs; Suri includes costs of other inputs, which she denotes X_{jit}^{H} and X_{jit}^{N}. Note also that Suri assumes (i) that "the quantity of seed used for a given area of land is the same whether it is hybrid or non-hybrid seed" (p.175) and (ii) the seed price is fixed across space (hence she uses b_t rather

than b_{it}).[10]

Now, denoting the optimised production as Y_{it}^{*H} and Y_{it}^{*N}, and optimal seed choice as s_{it}^{*}, the farm chooses to plant hybrid seed when:

$$Y_{it}^{*H} - Y_{it}^{*N} > \frac{a_{it} + b_{it} \cdot s_{it}^{*}}{p_{it}}. \tag{3.3}$$

The key point here, of course, quite intuitive: the farm will adopt hybrid seed where (i) the farm's *individual* return to hybrid seed is higher, and/or (ii) the farm's *individual* cost of obtaining hybrid seed is lower.

3.2.2 Production functions

Having imposed that optimal seed choice does not vary between hybrid and non-hybrid seed, Suri can abstract away from seed quantity in the production functions. Using Cobb-Douglas functions, expressed in logs, we have:

$$y_{it}^{H} = \beta^{H} + u_{it}^{H}, \tag{3.4}$$
$$y_{it}^{N} = \beta^{N} + u_{it}^{N}. \tag{3.5}$$

β^{H} and β^{N} therefore represent the 'sector-specific aggregate returns to yields'; u_{it}^{H} and u_{it}^{N} are 'sector-specific errors that may be the composite of time-invariant farm characteristics and time-varying shocks to production'.[11]

We then divide the error term into precisely those time-invariant farm characteristics and the time-varying shocks:

$$u_{it}^{H} = \theta_{i}^{H} + \xi_{it}^{H} \tag{3.6}$$
$$u_{it}^{N} = \theta_{i}^{N} + \xi_{it}^{N}. \tag{3.7}$$

As Suri puts it, "[f]armers are assumed to know θ_{i}^{H} and θ_{i}^{N}, but not ξ_{it}^{H} and ξ_{it}^{N}, when making their seed choice" (p.177). She discusses this assumption in more detail in her section 4.4 (in which, for example, she explains that the assumption essentially implies that "the unobserved time-varying variables that drive the switching should not be correlated with yields": p.183.)

[10] On the first assumption — about the quantity of seed used — Suri says (in her footnote 24), "There are standard seeding rates for maize that do not vary by seed type. This is borne out by the empirical seeding rates not varying... across hybrid and non-hybrid sectors; see Tables IIA and IIB." Query, however, whether this has implications for the production function; for example, does this imply that the marginal returns to quantity of hybrid seed are equivalent to the marginal returns to the quantity of non-hybrid seed?

[11] Note that Suri denotes β_{t}^{H} and β_{t}^{N}, but later (on page 184) assumes $\beta_{t}^{H} - \beta_{t}^{N} \equiv \beta \; \forall \; t$. I therefore remove the time subscript from the outset.

θ_i^H and θ_i^N refer to farm i's productivity in hybrid and non-hybrid seed respectively. When we turn to the empirical estimation, we will want to learn how important are productivity differences — that is, *relative* differences in productivity — in explaining hybrid seed adoption. Therefore, we express both θ_i^H and θ_i^N as a linear projection on the *relative productivity advantage*, $\theta_i^H - \theta_i^N$:

$$\theta_i^H = b_H \left(\theta_i^H - \theta_i^N\right) + \tau_i; \tag{3.8}$$

$$\theta_i^N = b_N \left(\theta_i^H - \theta_i^N\right) + \tau_i. \tag{3.9}$$

We can therefore interpret b_H as the *average* relationship between *relative productivity advantage* and *actual productivity in hybrid seed*; the residual, τ_i, is the farm's *absolute advantage*. Similarly, we interpret b_N as the *average* relationship between *relative productivity advantage* and *actual productivity in non-hybrid seed*. Note that $\mathbb{E}(\tau_i \mid \theta_i^H - \theta_i^N) = 0$ (because τ_i is the residual from a linear projection), and that the τ_i is the same term in equations 3.8 and 3.9.

3.2.3 Comparative advantage and absolute advantage

Suri discusses comparative and absolute advantage in terms of a series of linear projections; this follows closely the work of Lemieux (1998) (see also Card (1996)). We can define farmer-specific comparative advantage as:

$$\theta_i \equiv b_N \left(\theta_i^H - \theta_i^N\right). \tag{3.10}$$

The percentage difference between b_H and b_N is therefore:

$$\phi \equiv \frac{b_H}{b_N} - 1. \tag{3.11}$$

Then we have:

$$\theta_i^H = (\phi + 1) \cdot \theta_i + \tau_i; \tag{3.12}$$

$$\theta_i^N = \theta_i + \tau_i, \tag{3.13}$$

implying that, by definition:

$$\theta_i^H - \theta_i^N \equiv \phi \cdot \theta_i. \tag{3.14}$$

A simple example: For illustrative purposes, let's consider a simple case in which the population distribution of (θ_i^H, θ_i^N) is bivariate normal with a covariance of 2:

$$\begin{pmatrix} \theta_i^H \\ \theta_i^N \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}\right). \tag{3.15}$$

Figure 3.3 illustrates the joint distribution of (θ_i^N, θ_i^H) in this example. Figure 3.4 shows projections on $\theta_i^H - \theta_i^N$ of θ_i^H and θ_i^N respectively; in this example, we obtain $b_H = 1.5$ and $b_N = 0.5$. The fitted values from the righthand graph define θ_i.

Figure 3.3: **A simple example: Joint distribution of** (θ_i^N, θ_i^H)

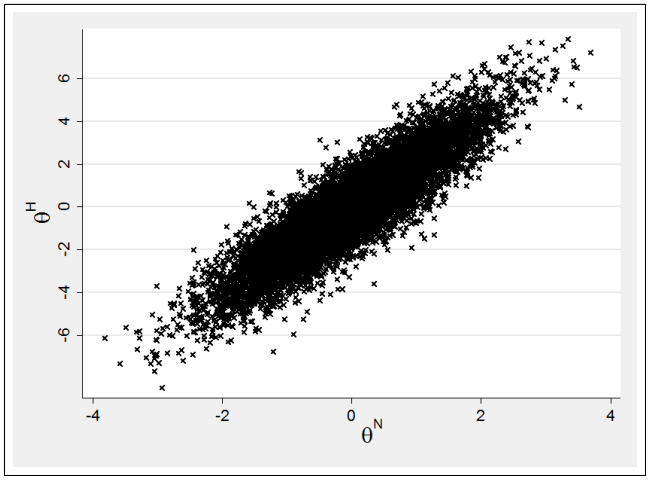
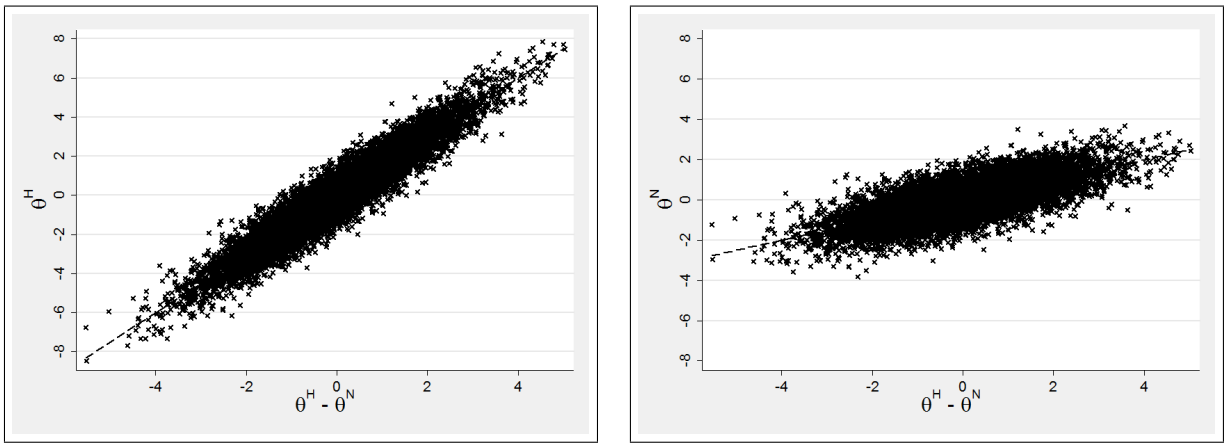


Figure 3.4: **A simple example: The role of** b_H **and** b_N

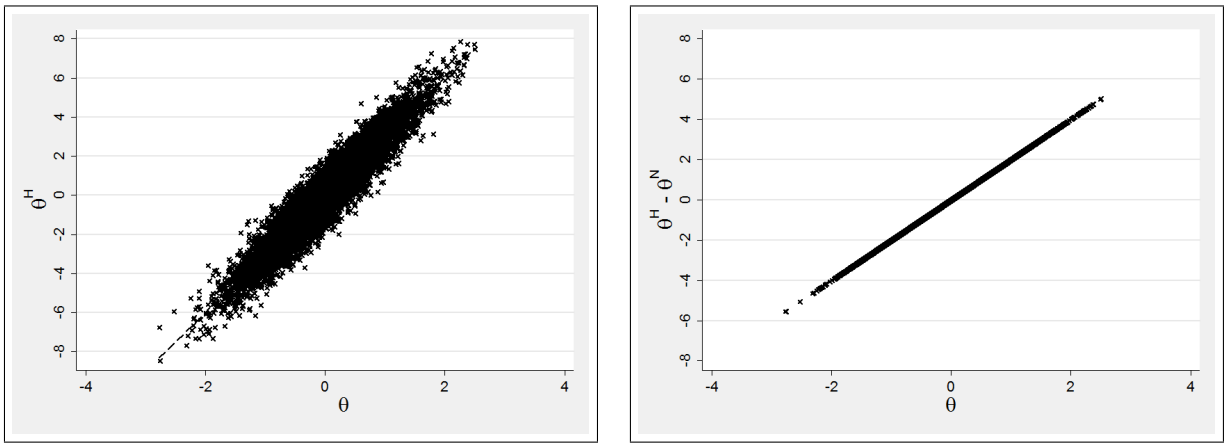


The left graph in Figure 3.5 shows a linear projection of θ_i^H on θ_i; the slope is $\phi + 1 = 3$, implying that $\phi = 2$. Alternatively, consider the right graph; this shows the identity $\theta_i^H - \theta_i^N \equiv \phi \cdot \theta_i$. Suri writes (p.178) that:

> θ_i measures farmer i's relative productivity in hybrid over nonhybrid, that is, his comparative advantage in hybrid.

(My personal interpretation is that this statement is true *if* $b_N > 0$. My interpretation is that, if $b_N < 0$, θ_i effectively measures relative productivity in *nonhybrid over hybrid*.)

Figure 3.5: **A simple example: The role of ϕ**



In contrast, the parameter ϕ captures the relative importance of θ_i for the farm's hybrid seed productivity, θ_i^H. As Suri explains (page 178):

> The coefficient ϕ therefore describes the sorting in the economy. If $\phi < 0$, there is less inequality in yields in this economy as compared to an economy where individuals are randomly allocated to a technology. On the other hand, if $\phi > 0$, then the self-selection process leads to greater inequality in yields.

Back to the production functions: We can therefore rewrite the production functions as:

$$y_{it}^H = \beta^H + \tau_i + (\phi + 1) \cdot \theta_i + \xi_{it}^H, \tag{3.16}$$
$$y_{it}^N = \beta^N + \tau_i + \theta_i + \xi_{it}^N. \tag{3.17}$$

Then we can define the dummy variable h_{it} as an indicator for whether the farm uses hybrid seed, so we have:

$$y_{it} = h_{it} \cdot y_{it}^H + (1 - h_{it}) \cdot y_{it}^N \tag{3.18}$$
$$= \beta^N + \theta_i + \left(\beta^H - \beta^N\right) \cdot h_{it} + \phi\theta_i h_{it} + \tau_i + \varepsilon_{it} \tag{3.19}$$
$$= \delta + \beta h_{it} + \theta_i + \phi\theta_i h_{it} + u_{it}. \tag{3.20}$$

3.3 Empirical method: The Correlated Random Coefficient model

Suri uses a Correlated Random Coefficient model in order to identify the structural objects of interest. You may be wonder why a simple Randomised Controlled Trial could not be used; Suri's footnote 5 is an interesting response to this:

> The questions posed here cannot be answered with an experiment which randomizes the technology across farmers without specific assumptions. With experimental data, one can test for the presence of heterogeneous returns, but estimating the distribution of returns requires assumptions about the underlying selection process, which is randomized away in such an experiment...

Suri uses a two-period panel dataset — with observations in 1997 and again in 2004. Thus, there are four possible combinations of hybrid seed adoption, as Table 3.1 shows.

Equation 3.20 is a '*correlated random coefficient*' specification — in that farm heterogeneity θ_i is allowed not merely to enter as a separate term (as it would in a fixed-effect model) but also enters the *coefficient* on hybrid adoption, $(\beta + \phi\theta_i)$. Suri presents a general discussion of how equation 3.20 may be estimated — including the case of covariates — but we will take a more intuitive approach in these notes.

Table 3.1: **Seed histories**

Seed Decision			
1997	2004	Suri's term	Our notation (averages)
non-hybrid	non-hybrid	'**non-hybrid stayers**'	y_{00}; θ_{00}
non-hybrid	hybrid	'**joiners**'	y_{01}; θ_{01}
hybrid	non-hybrid	'**leavers**'	y_{10}; θ_{10}
hybrid	hybrid	'**hybrid stayers**'	y_{11}; θ_{11}

Specifically, we will consider taking the mean production for each group in Table 3.1; in the case of 'leavers' and 'joiners', we will take means separately by time period. Each mean can be interpreted using equation 3.20. Note that our earlier assumption about the role of time-varying unobservables implies that:

$$\mathbb{E}(u_{it} \mid \theta_i, h_{i1}, h_{i2}) = 0. \tag{3.21}$$

First, consider the 'non-hybrid' group; for this group, we obtain:

$$y_{00,1997} = y_{00,2004} = \delta + \theta_{00} \qquad \text{for 1997 and 2004.} \tag{3.22}$$

Symmetrically, for the 'hybrid' group, we obtain:

$$y_{11,1997} = y_{11,2004} = \delta + \beta + \theta_{11} + \phi\theta_{11} \qquad \text{for 1997 and 2004.} \tag{3.23}$$

For the 'joiners' group, we obtain:

$$y_{01,1997} = \delta + \theta_{01} \qquad \text{for 1997, and} \tag{3.24}$$

$$y_{01,2004} = \delta + \beta + \theta_{01} + \phi\theta_{01} \qquad \text{for 2004.} \tag{3.25}$$

simon.quinn@economics.ox.ac.uk

Finally, for the 'leavers' group, we obtain:

$$y_{10,1997} = \delta + \beta + \theta_{10} + \phi\theta_{10} \quad \text{for 1997, and} \tag{3.26}$$

$$y_{10,2004} = \delta + \theta_{10} \quad \text{for 2004.} \tag{3.27}$$

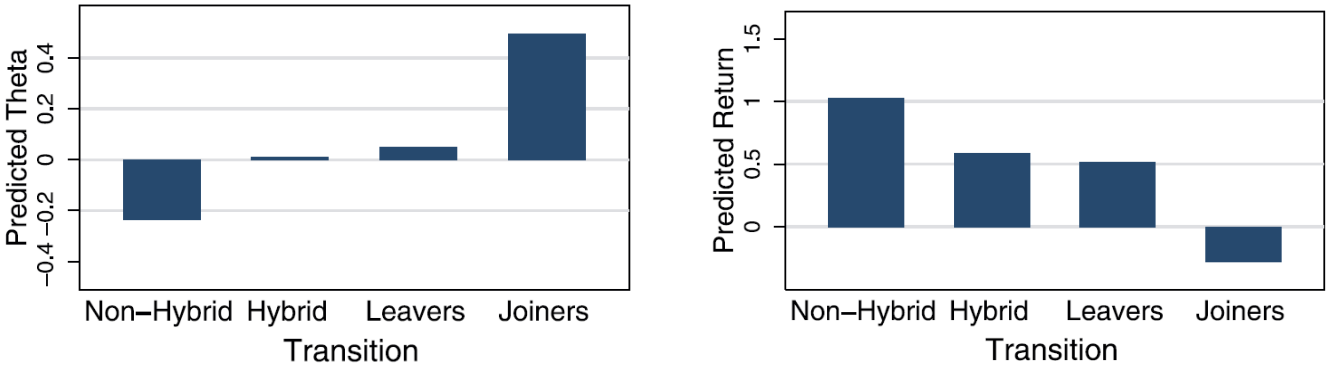
Additionally, Suri normalises θ_i such that $\sum_{i=1}^{N} \theta_i = 0$; this implies that we can obtain θ_{00} from the estimates of θ_{01}, θ_{10} and θ_{11}. Therefore, we have *five* 'structural parameters' of interest: θ_{01}, θ_{10} and θ_{11}, β and ϕ. But the preceding discussion shows that we can identify *six* linear combinations.[12] The system is therefore *over-identified*.

We will not discuss the actual estimation in any detail; in short, Suri uses a 'minimum distance' method to fit the model efficiently. She also reports a χ^2 over-identification statistic.[13]

3.4 Empirical results: Hybrid maize adoption among Kenyan farmers

For our purposes, Suri's main results are reported in her Table VIIIA, on page 196. I leave this for you to consider in detail. Figure 3.6 summarises the key results; the left figure (Suri's Figure 5A) shows the means of $\hat{\theta}_i$, and the right figure (Suri's Figure 5B) shows $\hat{\beta} + \hat{\phi} \cdot \hat{\theta}_i$.

Figure 3.6: **Suri's estimates of heterogeneous returns**



[12] There is, however, one important case in which the six linear combinations collapse to four, and the parameter ϕ is then not identified. Can you find this case? Suri discusses the issue on page 186 of her paper.

[13] Do you notice anything interesting about the magnitude of the χ^2 statistics reported?

Suri estimates $\phi < 0$, implying that — as she explains on page 194 —

> ... selection into hybrid is negative, with the farmers having the lowest yields in nonhybrid having the highest returns to planting hybrid. The estimated ϕ is consistently negative, which illustrates that the households that do better on average, do relatively worse at hybrid as the sign of ϕ describes the sorting process.

When you read Suri's paper — and, in particular, when you read her 'discussion' section — you may wish to consider the following questions:

(i) Suri finds "very large counterfactual returns to growing hybrid for the nonhybrid stayers" (p.200); why, then, do those farmers not adopt the technology?

(ii) Does Suri find any evidence for the Griliches-style logistic-shaped adoption curve?

(iii) Suri argues that credit constraints "do not seem to be of first order importance"; but how, in principle, *might* such constraints explain the patterns in Figure 3.6?

(iv) Suri concludes by asserting (p.205) that "the heterogeneity in returns to the hybrid technology, on the whole, suggests quite rational and relatively unconstrained adoption of existing hybrid strains, in contrast to the evidence from the experimental and IV literatures". These other literatures include, for example, Duflo *et al* (2010), who argue that — in the Kenyan context — "small, time-limited discounts can potentially help present-biased farmers commit to fertilizer use, and thus overcome procrastination problems..." (p.42). Can these two results be reconciled?

(v) Michler et al (2018) estimate returns to adoption of improved chickpea in Ethiopia; they argue that "economic measures of returns may be more relevant than increases in yields in explaining technology adoption decisions". What do they mean? Should their results change our interpretation of Suri (2011)?

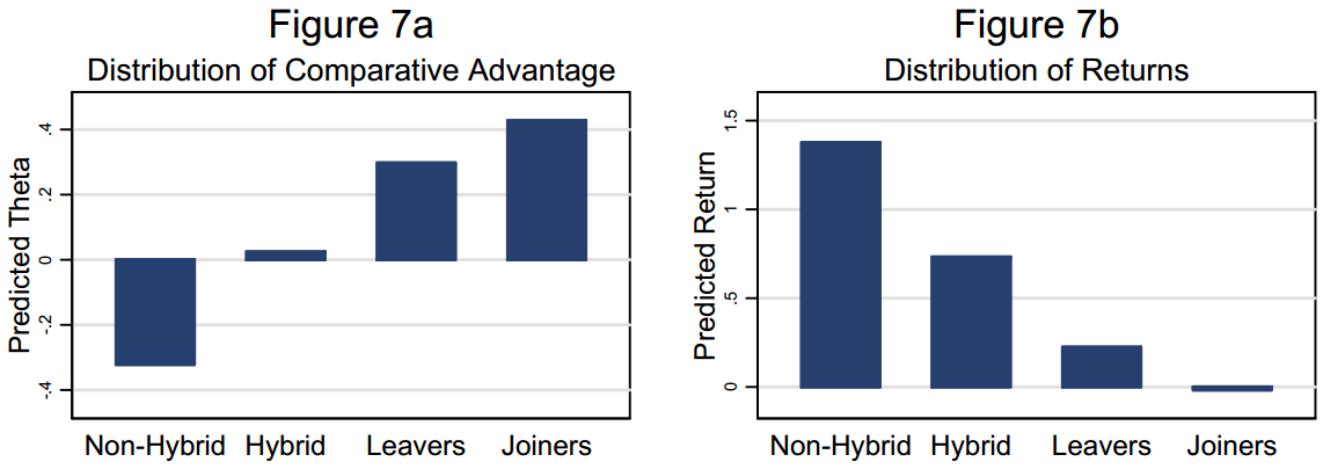
You may wish to note the following points as you work through Suri's analysis.

(i) As already noted, it seems logical to interpret θ_i as relative productivity in nonhybrid over hybrid, not the converse.

(ii) It is possible that the estimate of λ_3 in the first column of Table VIIIA should be '-1.636', rather than 1.636.

(iii) Suri repeatedly refers to some farmers having returns of approximately 150%. She also notes several times that 'marginal farmers' (*i.e.* joiners and leavers) have returns close to zero. Both of these claims may make more sense when referring to Figure 7b in an earlier version of the paper (namely, Suri (2006)):

Figure 3.7: **Figures 7a and 7b from Suri (2006)**



4 Lecture 4: Firm Size

References:

- [⋆] Amirapu, A., and Gechter, M. (2017): "Labor Regulations and the Cost of Corruption: Evidence from the Indian Firm Size Distribution," *Review of Economics and Statistics*, 102(1), 34–48.
- [⋆] Söderbom, M., and Teal, F. (2004): "Size and Efficiency in African Manufacturing Firms: Evidence from Firm-Level Panel Data," *Journal of Development Economics*, 73(1), 369–394.
- Akcigit, U., Alp, H., and Peters, M. (2020): "Lack of Selection and Limits to Delegation: Firm Dynamics in Developing Countries," *American Economic Review*, 111(1), 231–275.
- Ceci-Renaud, N., and Chevalier, P. (2010): "L'impact des seuils de 10, 20 et 50 salariés sur la taille des enterprises francáises," *Économic et Statistique*, 437:29–45.
- Garicano, L., Lelarge, C., and Van Reenen, J. (2016): "Firm Size Distortions and the Productivity Distribution: Evidence from France," *American Economic Review*, 106(11), 3439–3479. (Online appendix available at `https://assets.aeaweb.org/assets/production/files/2511.pdf`.)
- Gourio, F., and Roys, N. (2014): "Size-Dependent Regulations, Firm Size Distribution, and Reallocation," *Quantitative Economics*, 5(2):377-0416.
- Lucas, R. (1978): "On the Size Distribution of Business Firms," *The Bell Journal of Economics*, 9(2), 508–523.
- Rauch, J.E. (1991): "Modelling the Informal Sector Formally," *Journal of Development Economics*, 35(1), 33–47.
- Sutton, J. (1997): "Gibrat's Legacy," *Journal of Economic Literature*, 35(1), 40–59.

4.1 Introduction

4.1.1 Heterogeneity in constraints and in performance

How do large firms' constraints differ from those of small firms? Until now, we have hardly discussed the issue of firm size. To be sure, firm size has *entered* our analysis, but it has only done so incidentally — for example, as an incidental part of a story about returns to capital, or as an aspect of a firm's optimisation problem. In this sense, the issue of firm size and firm expansion has been a bit like the proverbial 'elephant in the room' — an important issue, but not one that we have really wanted to discuss.[14]

[14] The metaphorical 'elephant in the room' is, of course, not to be confused with the '800 pound gorilla in the room'.

Yet important the issue is: it informs our understanding of *heterogeneity in firms' constraints*, and *heterogeneity in firms' performance*. Our discussion of 'firms and accumulation' in Lecture 1 was clearly directed towards understanding a very different *kind* of firm than that considered in Lecture 2 (on 'firms and production'). In the same way that physicists may worry about applying the same theories to particles of very different sizes, development economists may feel that the forces acting upon very small firms are fundamentally different to those faced by much larger enterprises. Our goal in this lecture is certainly not to provide any 'Grand Unified Theory' for different types of firms — however, we ought to have some idea of why different firms may operate at different sizes, and of the different constraints that such firms can face.

In this lecture, we will do three things:

(i) We will consider a formal model in which regulation and heterogeneity in management quality combine to generate a 'broken power law' distribution of firm sizes;

(ii) We will use predictions from that model to draw empirical conclusions about the costs of regulation; and

(iii) Building on the methods that we discussed in Lecture 2, we will discuss empirical results on the relationship between firm size and factor costs.

4.2 Theoretical model: Regulation, management quality and firm size

In this lecture, we will primarily consider recent work by Amirapu and Gechter (2017), which uses the distribution of firm sizes in India to learn about the cost of firm regulation and the role of corruption. We will say little in this lecture about corruption; I leave you to read this in Amirapu and Gechter's paper. Instead, we will focus on the authors' theoretical model, and on their basic method for analysing the firm size distribution.

4.2.1 Enterprise sizes in India

At its core, Amirapu and Gechter's theoretical analysis rests upon power law distributions: namely, a power law distribution in management quality, generating a power law distribution in firm size. Figure 4.9 replicates Figure 2 from Amirapu and Gechter's paper; it shows the density of enterprises of different sizes in India, on a log-log scale. Note that this is (approximately) linear; this is indicative of a power law in firm size. As Amirapu and Gechter note, this kind of power-law distribution has been observed in firm-size distributions across a number of countries (both developed and developing). Empirically, the power-law distribution is similar to the log-normal. The log-normal has also been used effectively to model firm size distributions — following, in particular, the seminal work of the French engineer Robert Gibrat, analysing the size distribution of French manufacturing firms using data from the early 1920s: see Sutton (1997).

Figure 4.1: **'2005 Log-Log Distribution of Establishment Size': Figure 2 in Amirapu and Gechter (2017)**

Note: Both axes are on a log scale. Total number of workers is the number of workers usually working daily in an establishment. Source: 2005 Economic Census of India.

You should also notice the 'break' in the distribution between enterprises having nine and having ten employees: this is central to the authors' analysis, and reflects a critical size for the application of a wide range of Indian regulatory standards. As the authors explain (p.9):

> The major regulations that start to apply once an establishment employs 10 or more workers include the following: establishments must register with the government, meet various workplace safety requirements (under the Factories Act for manufacturing establishments that use power and The Building and Other Construction Workers' Act for construction-related establishments, for example), pay insurance/social security taxes (under the Employees' State Insurance Act), distribute gratuities (under the Payment of Gratuity Act) and they must bear a greater administrative burden (under, for example, the Labor Laws Act). Other regulations are indirectly size-based, because they reference laws with size-based aspects. For example, the Maternity Benefits Act only applies to establishments designated as "factories" under the Factories Act, which means it only applies to establishments with more than 10 workers.

4.2.2 A model of optimal firm size under regulation

Amirapu and Gechter use an elegant theoretical model that builds on the earlier work of Garicano, Le Large and Van Reenen (2016); in turn, this builds on earlier work by Lucas (1978), Rauch (1991), Ceci-Renaud and Chevalier (2010), and Gourio and Roys (2014).

In my view, this model is useful for three reasons. First, it considers whether, in theory, we should expect firms of different sizes to co-exist; this essentially implements the seminal work of Lucas (1978), in which different firm sizes exist in equilibrium through heterogeneity in management quality. Second, the model allows directly for differential regulation based upon a critical point in the firm distribution. Third, it is a 'structural model': a model whose parameters can be estimated directly from data.

4.2.3 The role of management quality

Fundamentally, the Lucas (1978) approach — and the literature that follows it — attributes heterogeneity in firm size to heterogeneity in management quality. We therefore begin by specifying firm production, using a simple model that combines labour with managerial ability.

Assumption 8 (Production process) *Each firm has a single manager with talent α, and hires $n(\alpha)$ employees (where $n(\alpha) \in \mathbb{R}^+ \ \forall \ \alpha$). At a given wage rate w, firm output Q is determined by Cobb-Douglas production, where α plays the role of firm technology:*

$$Q(x, N; w) = \alpha \cdot n^{\theta}. \tag{4.1}$$

Assumption 9 (Strictly diminishing returns) *We assume that the manager has a limited 'span of control', so that there are strictly diminishing returns to labour input:*

$$\theta \in (0, 1). \tag{4.2}$$

It follows that a firm with management quality α earns profit of:

$$\pi(\alpha, n; w) \equiv \alpha \cdot n^{\theta} - w \cdot n. \tag{4.3}$$

Thus, the manager will choose labour input so that:

$$\left.\frac{\partial \pi(\alpha, n; w)}{\partial n}\right|_{n=n^*(\alpha;w)} = 0 \tag{4.4}$$

$$\Leftrightarrow n^*(\alpha; w) = \left(\frac{\theta\alpha}{w}\right)^{\gamma}, \tag{4.5}$$

where, for convenience, we use $\gamma \equiv \dfrac{1}{1-\theta} > 1$. Therefore, a firm hires more labour if the manager is more talented; conversely, the firm hires less labour if the wage is higher.

Assumption 10 (Form of government regulation) *Regulation occurs* solely *through a per-worker tax on labour, of $\tau \cdot w$. Firms must pay this tax if they have $n > N$. Firms with size $n \leq N$ do not pay the tax.*

To solve the model, we need to consider two possible regimes: regulated, and unregulated. We calculate optimal profits for any given firm under each of those two regimes, then calculate which regime is optimal.

Profits in *regulated* firms: In regulated firms, the cost to the firm of each additional worker is $(1+\tau)\cdot w$. Therefore, a regulated firm with management quality α will hire $n^*(\alpha;(1+\tau)\cdot w)$, and profits in regulated firms will be:

$$\pi^R(\alpha) \equiv \alpha \cdot \left(\frac{\theta\alpha}{(1+\tau)\cdot w}\right)^{\theta\gamma} - (1+\tau)\cdot w \cdot \left(\frac{\theta\alpha}{(1+\tau)\cdot w}\right)^{\gamma}. \tag{4.6}$$

Profits in *unregulated* firms: Define by α_1 *the management quality that would, in the absence of any regulation, choose a firm size of* N. That is,

$$N \equiv \left(\frac{\theta\alpha_1}{w}\right)^{\gamma} \tag{4.7}$$

$$\Leftrightarrow \alpha_1 = \left(\frac{w}{\theta}\right)\cdot N^{1-\theta}. \tag{4.8}$$

In that case, an unregulated firm — that is, a firm choosing $n \leq N$ — will adopt a cutoff strategy: *choose* $n < N$ *for* $\alpha < \alpha_1$*; choose* $n = N$ *for* $\alpha \geq \overline{\alpha_1}$. Or, formally:

$$n = \begin{cases} \left(\frac{\theta\alpha}{w}\right)^{\gamma} & \text{for } \alpha < \alpha_1; \\ \\ N & \text{for } \alpha \geq \alpha_1. \end{cases} \tag{4.9}$$

Profit for an unregulated firm therefore also follows a piecewise function:

$$\pi^U(\alpha) = \begin{cases} \alpha \cdot \left(\frac{\theta\alpha}{w}\right)^{\theta\gamma} - w \cdot \left(\frac{\theta\alpha}{w}\right)^{\gamma} & \text{for } \alpha < \alpha_1; \\ \\ \alpha \cdot N^{\theta} - w \cdot N & \text{for } \alpha \geq \alpha_1. \end{cases} \tag{4.10}$$

Comparing rents across sectors: Figure 4.2 shows the functions $r^U(x)$ and $r^R(x)$; this is essentially Rauch's Figure 1 (page 38). Note the use of the cutoff α_2, defined such that $\pi^U(\alpha_2) \equiv \pi^R(\alpha_2)$; further, note that we assume $\alpha_2 > \alpha_1$, implying that the regulation actually binds for at least some firms.

Implications for firm size: Assuming that each firm chooses optimally whether to be regulated or unregulated, we can now predict firm size as a function of managerial ability:

$$n(\alpha) = \begin{cases} \left(\frac{\theta\alpha}{w}\right)^{\gamma} & \text{if } \alpha \in [\underline{\alpha}, \alpha_1]; \\ N & \text{if } \alpha \in (\alpha_1, \alpha_2); \\ \left(\frac{\theta\alpha}{(1+\tau)\cdot w}\right)^{\gamma} & \text{if } \alpha \geq \alpha_2. \end{cases} \tag{4.11}$$

Figure 4.2: **Entrepreneurial rent across two sectors**

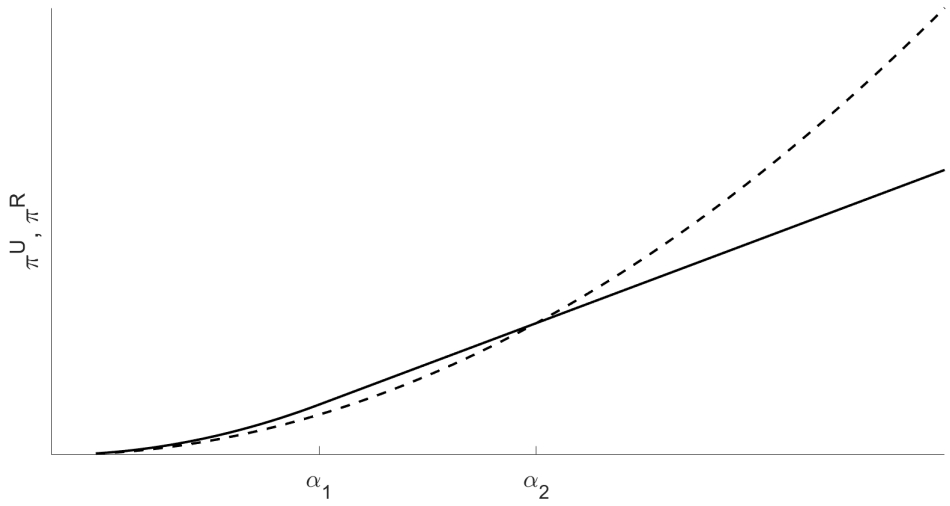
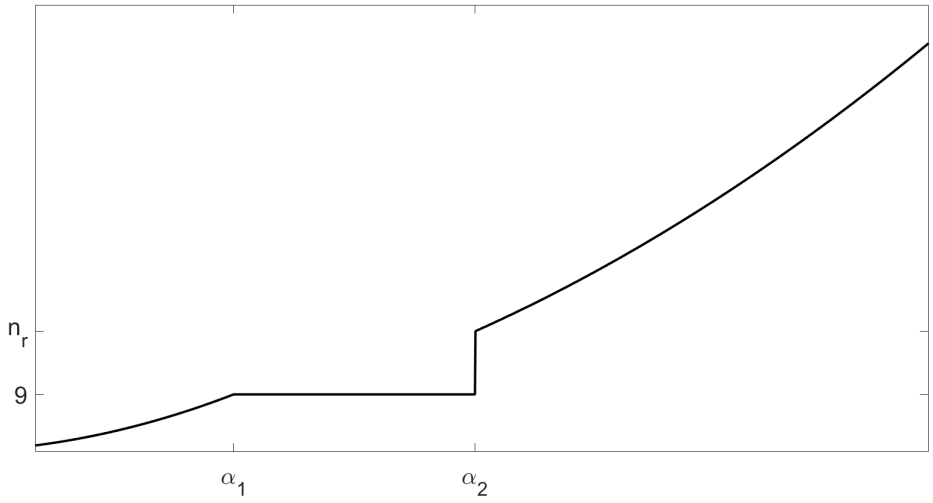


Figure 4.3 illustrates. Note that the smallest firm of a size larger than N has size n_r, where:

$$n_r = \left(\frac{\theta \alpha_2}{(1+\tau) \cdot w} \right)^{\gamma}. \tag{4.12}$$

That is, if there were no regulation, a firm having management quality α_2 would optimally choose size $n_r \cdot (1+\tau)^{\gamma}$. (Note that Garicano et al refer to n_r as the "upper employment threshold'.)

Figure 4.3: **Optimal firm size and management quality**



4.3 Identifying the cost of regulation

4.3.1 An aside: Microeconomic theory meets quasi-experimental econometrics...

We will turn to consider empirical analysis shortly. Before we do so, suppose that a researcher reasons as follows:

> I would like to measure the effect of the increased regulatory burden upon firm performance. To do this, I will compare the outcomes for firms *just above* the cutoff N with those *just below*; that is, I will use a *regression discontinuity design*.

Why would this *not* be a valid strategy for measuring the effect of the minimum wage? Can you think of any simple empirical test to show that this is not a valid strategy in this kind of context?

4.3.2 Distribution of management quality

First, let's make an assumption about the distribution of management quality in the population. Let the minimum possible management quality be $\underline{\alpha}$, and define:

$$\beta_\alpha = \frac{\beta - \theta}{1 - \theta}, \tag{4.13}$$

where $\beta > 1$.

With these definitions in hand, we assume that α has a power law distribution (denoting the *cdf* by Φ and the *pdf* by ϕ):

$$\begin{aligned} \Phi(\alpha) &= 1 - \left(\frac{\alpha}{\underline{\alpha}}\right)^{1-\beta_\alpha}; \\ \therefore \phi(\alpha) &= c_\alpha \cdot \alpha^{-\beta_\alpha}, \\ \text{where } c_\alpha &= (\beta_\alpha - 1) \cdot \underline{\alpha}^{\beta_\alpha - 1}. \end{aligned}$$

For simplicity, let's begin by considering firms choosing $n < N$. For these firms, we have:

$$n(\alpha) = \left(\frac{\theta\alpha}{w}\right)^{\gamma}; \tag{4.14}$$

$$\therefore \alpha(n) = \frac{w \cdot n^{1-\theta}}{\theta}; \tag{4.15}$$

$$\therefore \frac{d\alpha}{dn} = (1 - \theta) \cdot \frac{w}{\theta} \cdot n^{-\theta}. \tag{4.16}$$

Therefore, using the change of variables formula, we can say:

$$\chi(n) = \left|\frac{d\alpha(n)}{dn}\right| \cdot \phi\left[\alpha(n)\right] \tag{4.17}$$

$$= (1 - \theta) \cdot \frac{w}{\theta} \cdot n^{-\theta} \cdot c_\alpha \cdot \left(\frac{w}{\theta}\right)^{-\beta_\alpha} \cdot n^{-\beta_\alpha(1-\theta)} \tag{4.18}$$

$$= c_\alpha \cdot (1 - \theta) \cdot \left(\frac{\theta}{w}\right)^{\beta_\alpha - 1} \cdot n^{-\beta_\alpha(1-\theta)-\theta} \tag{4.19}$$

$$= c_\alpha \cdot (1 - \theta) \cdot \left(\frac{\theta}{w}\right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot n^{-\beta} \tag{4.20}$$

$$= C \cdot n^{-\beta}, \tag{4.21}$$

where we define $C \equiv c_\alpha \cdot (1 - \theta) \cdot \left(\frac{\theta}{w}\right)^{\left(\frac{\beta-1}{1-\theta}\right)}$ for simplicity.

Note that, under this structure, the power law in management ability translates into a power law in firm size:

$$\ln \chi(n) = \ln C - \beta \cdot \ln n. \tag{4.22}$$

For those firms choosing $n > N$, we can use exactly the same reasoning — substituting $(1+\tau) \cdot w$ in place of w. For those firms, we can therefore say:

$$\chi(n) = c_\alpha \cdot (1-\theta) \cdot \left(\frac{\theta}{(1+\tau) \cdot w} \right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot n^{-\beta} \tag{4.23}$$

$$= c_\alpha \cdot (1-\theta) \cdot \left(\frac{\theta}{w} \right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot n^{-\beta} \cdot (1+\tau)^{\left(\frac{1-\beta}{1-\theta}\right)}. \tag{4.24}$$

For simplicity, denote $T \equiv (1+\tau)^{\left(\frac{1-\beta}{1-\theta}\right)}$. Then, for firms choosing $n > N$, we can rewrite the earlier expression as:

$$\chi(n) = c_\alpha \cdot (1-\theta) \cdot \left(\frac{\theta}{w} \right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot n^{-\beta} \cdot T = C \cdot T \cdot n^{-\beta}. \tag{4.25}$$

4.3.3 Bunching at N

What is the mass of firms bunching at N? These are firms with $\alpha \in (\alpha_1, \alpha_2)$. Recall that a firm having $\alpha = \alpha_1$ optimally chooses $n = N$. Similarly, recall that — if there were no regulation — a firm having management quality α_2 would optimally choose size $n_r \cdot (1+\tau)^\gamma$. Therefore, we can calculate the mass of firms bunching by integrating over the firm size that bunching firms *would* choose, were they not facing regulation. That is:

$$\Pr(n = N) = \int_N^{[(1+\tau)^\gamma] \cdot n_r} c_\alpha \cdot (1-\theta) \cdot \left(\frac{\theta}{w} \right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot n^{-\beta} \, dn \tag{4.26}$$

$$= c_\alpha \cdot \frac{1-\theta}{1-\beta} \cdot \left(\frac{\theta}{w} \right)^{\left(\frac{\beta-1}{1-\theta}\right)} \cdot \left[(1+\tau)^{\gamma \cdot (1-\beta)} \cdot n_r^{1-\beta} - N^{1-\beta} \right] \tag{4.27}$$

$$= \frac{C}{\beta - 1} \cdot \left(N^{1-\beta} - T \cdot n_r^{1-\beta} \right), \tag{4.28}$$

where T and C are the same constants that we defined earlier.

4.3.4 Proportion of firms in each category

So far, we have discussed three categories — those firms bunching at N, those firms choosing $n < N$, and those firms choosing $n > N$. What proportion of firms lies in each category?

First, what proportion of firms choose $n < N$? We can solve this by integrating:

$$\Pr(n < N) = \frac{C}{1-\beta} \cdot \left(N^{1-\beta} - n_{\min}^{1-\beta} \right). \tag{4.29}$$

And what proportion of firms choose $n > N$? Again, by integrating, we have:

$$\Pr(n > N) = -\frac{CT}{1-\beta} \cdot n_r^{1-\beta}. \tag{4.30}$$

What about the intermediate category — the firms bunching at N? Well, trivially, these are all of the firms that are neither above nor below N. Therefore, we can find an alternative expression for δ:

$$\Pr(n = N) = 1 - \Pr(n < N) - \Pr(n > N) \tag{4.31}$$

$$= 1 - \frac{C}{\beta - 1} \cdot \left(n_{\min}^{1-\beta} - N^{1-\beta} + T \cdot n_r^{1-\beta} \right). \tag{4.32}$$

The term $n_{\min}$ has just entered for the first time. Fortunately, we can get rid of it just as quickly...

Assumption 11 (Continuity of earnings) *The worst manager of an existing firm is indifferent between managing that firm and being a wage worker:*

$$\pi^U(\alpha_{min}) = w \tag{4.33}$$

$$\Leftrightarrow \alpha_{min} \cdot n_{min}^{\theta} - w \cdot n_{min} = w. \tag{4.34}$$

Garicano et al describe this 'continuity of earnings' as arising from an important property of a larger solution concept — namely, the requirement that, in any solution, 'no agent wishes to change occupation (worker versus manager)' (p.3445).

Substituting from equation 4.15, we can say:

$$\frac{w}{\theta} \cdot n_{\min} - w \cdot n_{\min} = w, \tag{4.35}$$

$$\therefore n_{\min} = \frac{\theta}{1-\theta}. \tag{4.36}$$

Compare equations 4.28 and 4.32 — these equations provide two different expressions for the same object. Substituting in for $n_{\min}$ and equating these two expressions, we obtain:

$$\frac{C}{\beta - 1} \cdot \left(N^{1-\beta} - T \cdot n_r^{1-\beta} \right) = 1 - \frac{C}{\beta - 1} \cdot \left[\left(\frac{\theta}{1-\theta} \right)^{1-\beta} - N^{1-\beta} + T \cdot n_r^{1-\beta} \right], \tag{4.37}$$

implying that:

$$\frac{C}{\beta - 1} = \left(\frac{1-\theta}{\theta}\right)^{1-\beta}, \tag{4.38}$$

which itself then implies:

$$\Pr(n = N) = \left(\frac{1-\theta}{\theta}\right)^{1-\beta} \cdot \left(N^{1-\beta} - T \cdot n_r^{1-\beta}\right). \tag{4.39}$$

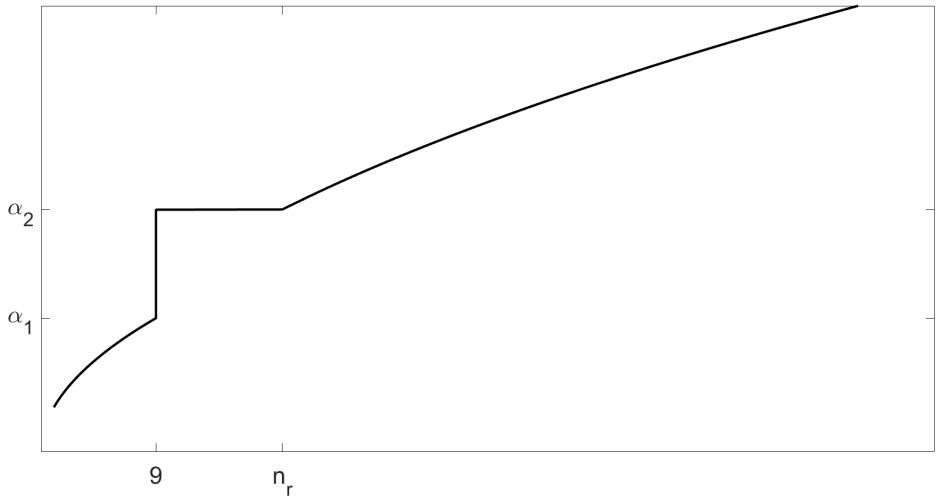
4.3.5 Density of firms across the distribution

Combining equations 4.21, 4.25 and 4.39, we can now write the density for firm size:

$$\chi(n) = \begin{cases} \left(\frac{1-\theta}{\theta}\right)^{1-\beta} \cdot (\beta - 1) \cdot n^{-\beta} & \text{if} \quad n \in \left[\frac{\theta}{1-\theta}, N\right); \\ \left(\frac{1-\theta}{\theta}\right)^{1-\beta} \cdot \left(N^{1-\beta} - T \cdot n_r^{1-\beta}\right) & \text{if} \quad n = N; \\ 0 & \text{if} \quad n \in \left(N, n_r\right); \\ \left(\frac{1-\theta}{\theta}\right)^{1-\beta} \cdot (\beta - 1) \cdot T n^{-\beta} & \text{if} \quad n \geq n_r. \end{cases} \tag{4.40}$$

To see the intuition of all these transformations, I think it is useful to consider a few graphs. First, in Figure 4.4, we have Figure 4.3, but flipped across the 45-degree line.

Figure 4.4: **Optimal firm size and management quality**



In Figure 4.5, we have the same graph, but plotted in log-log space. Figure 4.6 — not the most complicated graph you will ever see — shows the distribution of management

quality, in log-log space (where, for comparability with the previous graph, we have α on the y-axis). Finally, by combining the relationship in Figure 4.5 with the density in Figure 4.6, we obtain the broken power law of equation 4.40; this is illustrated in Figure 4.7.

Figure 4.5: **Optimal firm size and management quality: Log-log space**

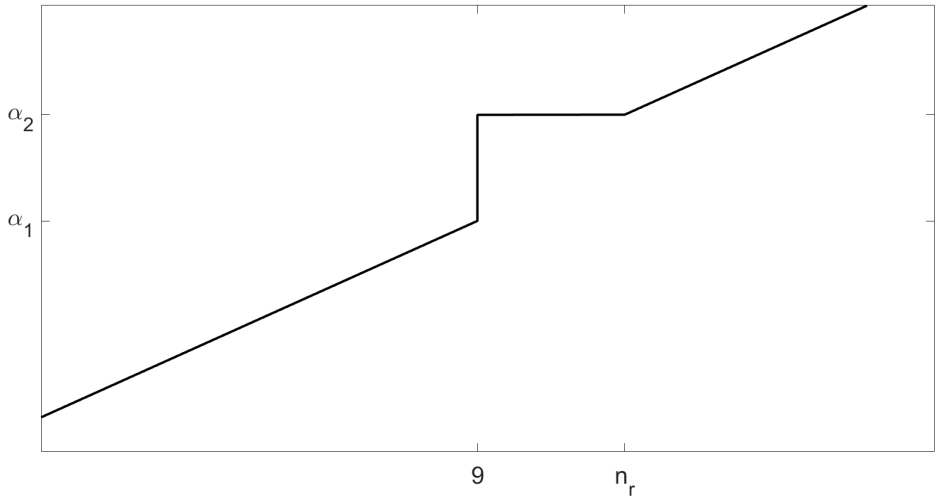


Figure 4.6: **Density of management quality: Log-log space**

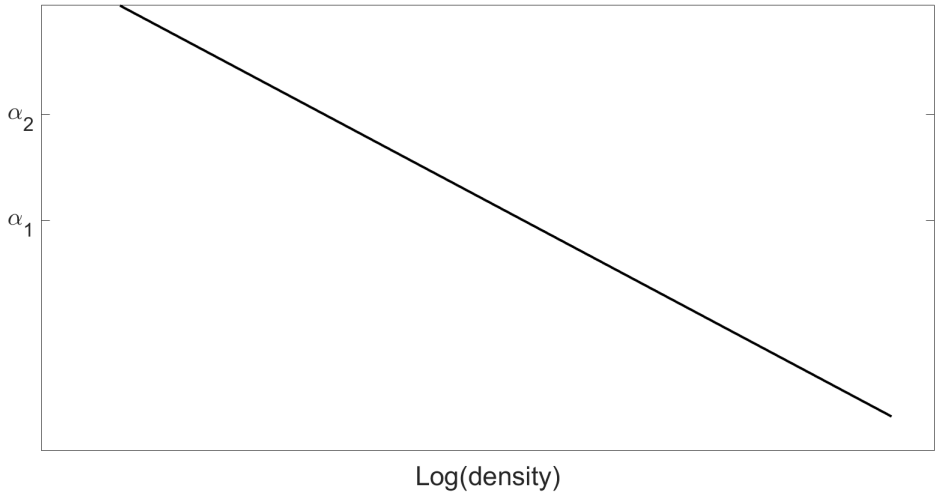
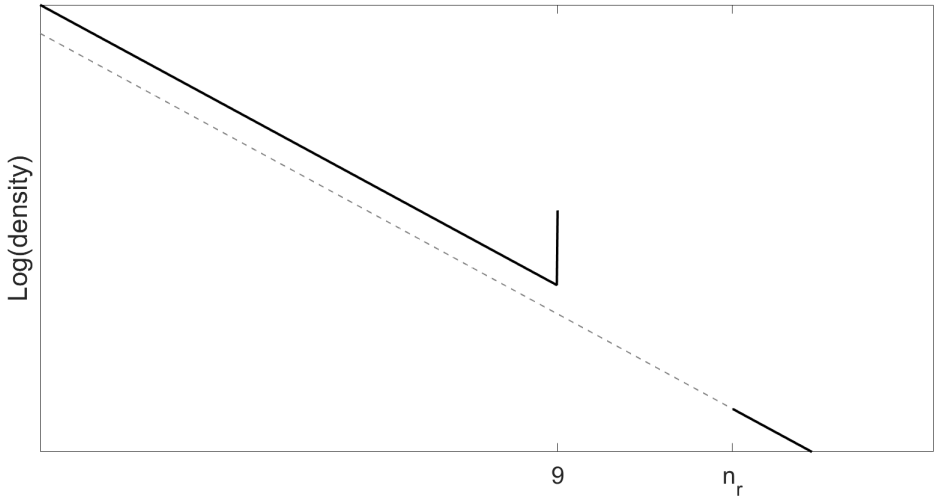
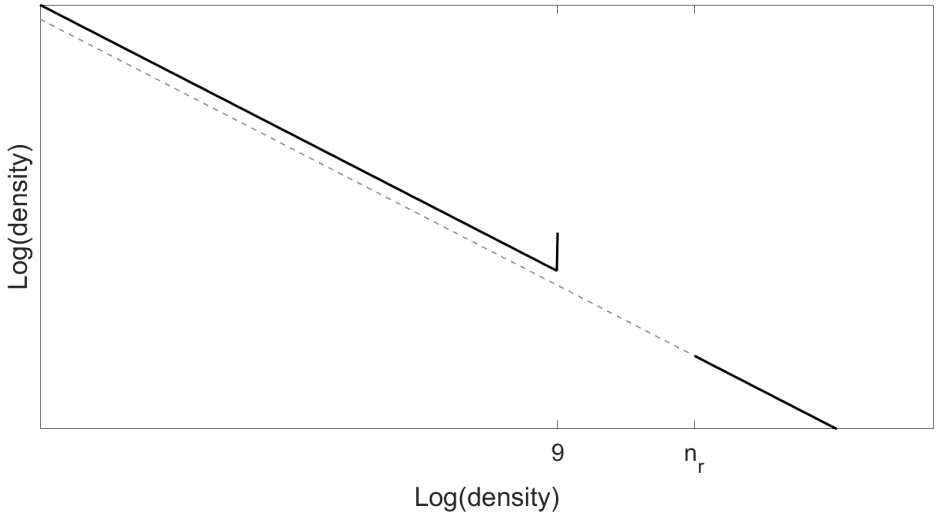


Figure 4.7: **A broken power law in firm size: Log-log space**



So what? Figure 4.7 — and, with it, equation 4.40 — solves what is sometimes known as the 'forward problem'; that is, it shows how a theoretical model makes predictions about data. This is a very important thing to be able to do. For example, in Figure 4.8, we consider what happens to our prediction about density when we halve τ.

Figure 4.8: **A broken power law in firm size, halving τ: Log-log space**



However, the true value of a structural model lies in going the other way — that is, taking data generated by a model and using that data to recover the underlying parameters. This is sometimes known as the 'inverse problem'. Intuitively, the inverse problem can be understood by asking a simple question: *'if I had an infinitely large dataset, generated by my model, would I be able to use that dataset to solve for the model parameters?'*. If the answer

is 'yes', we can say that the model is 'identified'.[15]

So, suppose that you had an infinitely large dataset of Indian firms, and suppose that the model considered in this lecture is literally the true data-generating process. That is, suppose that — without needing to worry about any specific estimation method — you could observe directly the density for each firm size. How would you use that data to solve for the parameters β, θ and τ? (*Hint*: Equation 4.40 is obviously highly relevant here...!)

4.4 Estimation and empirical results

Unfortunately, infinitely large datasets can take quite a long time to collect. In the real world, we cannot literally '*solve for*' the model parameters of interest; instead, we need to *estimate* those parameters, using the available data. Before estimating, we need to confront a troubling fact: our model predicts that there should be a 'hole' to the right of N; that is, there should be some firm sizes with *zero* density. Look again at Figure 4.9; you will see a clear break between the density at size nine and at size ten, but the density at size ten does not drop to zero. This is problematic. If we were to take equation 4.40 literally, the conclusion would be simple: the model fails. That is, the data includes some observations that the model declares to be completely impossible.[16]

Different authors take different approaches to this problem. For example, Garicano, Le Large and Van Reenen (2016) incorporate measurement error in employment. Gourio and Roys (2014) say that '[i]t would be incredible to attribute the presence of all these firms to measurement error' (p.392); those authors instead use a dynamic model, incorporating a sunk cost to be paid the first time that the firm crosses the employment threshold. In contrast, Amirapu and Gechter (2017) 'consider a fraction of firms to be inattentive to the threshold' (p.22); they assume that this fraction decreases towards zero as management quality increases.

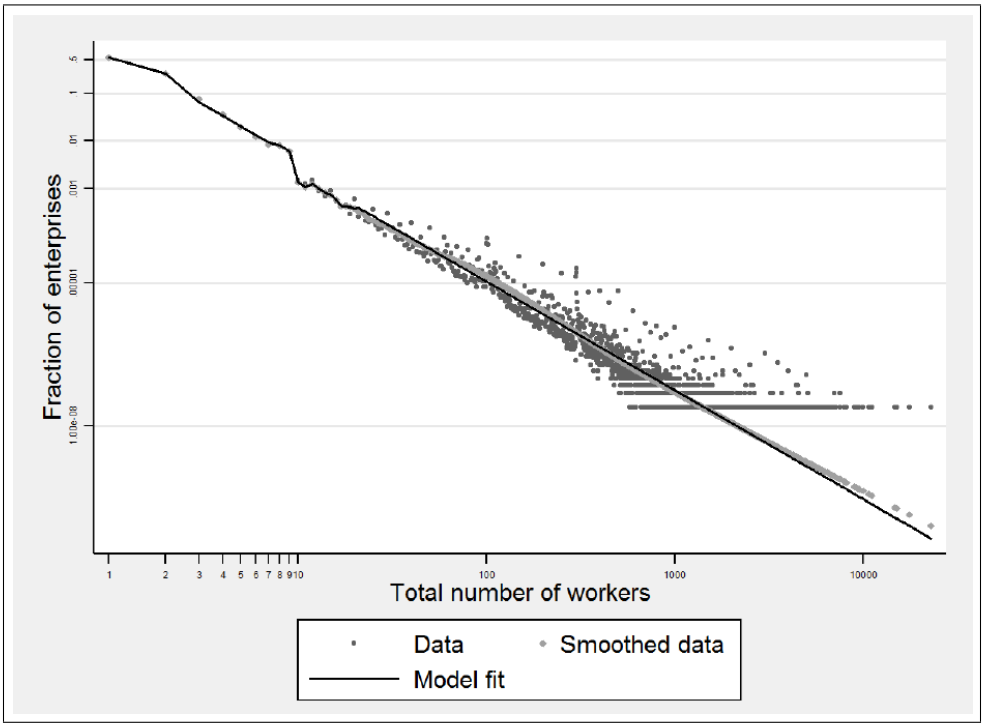
Having made some assumption to handle this problem, it is time to turn to estimation. Again, different authors have different approaches. Gourio and Roys (2014) use Indirect Inference to match model predictions against data, on both (i) 'the discontinuity in the firm size distribution', and 'the firm size distribution, conditional on having operated above 55 employees in the past' (p.379). Garicano, Le Large and Van Reenen (2016) estimate using Maximum Likelihood. Amirapu and Gechter (2017) use a non-parametric density estimator (having transformed their data to account for having a 'heavy-tailed' distribution); they estimate the density to the right of the critical firm size, then use this estimate to recover an estimate of τ. Figure 4.9 illustrates. I leave it to you to consider

[15] Look again at equation 2.11, in Lecture 2. In that case, the answer was 'no' — because, no matter how large the hypothetical dataset, there was no way of separating a_i from k_i^* or l_i^*.

[16] Thus, for example, if we were to estimate this model using Maximum Likelihood, we would find ourselves having to evaluate $\log(0)$. But the problem is more profound than just a computational error, of course...

the authors' empirical results (including the application of those results to consider the role of corruption).

Figure 4.9: **'Model Fit and Data': Figure 3 in Amirapu and Gechter (2017)**



4.5 Firm size and factor prices: Söderbom and Teal (2004)

We consided Söderbom and Teal (2004) in Lecture 2, as an example of using linear panel data to estimate the parameters of a Cobb-Douglas production function. To end this lecture, we now return to the same paper — particularly section 7 — to consider its results on how firms' constraints differ with firm size. I suggest the following questions for discussion.

(i) Find the first-order conditions for the following cost-minimisation problem:

$$\min_{K,L} \; r \cdot K + w \cdot L \tag{4.41}$$

$$\text{subject to } \; \overline{Y} = A \cdot K^{\alpha} L^{\beta}. \tag{4.42}$$

Do you obtain Söderbom and Teal's equation 3? How do the production function estimates (considered in Lecture 2) assist here?

(ii) Interpret the following figure (Figure 2 from Söderbom and Teal (2004)).

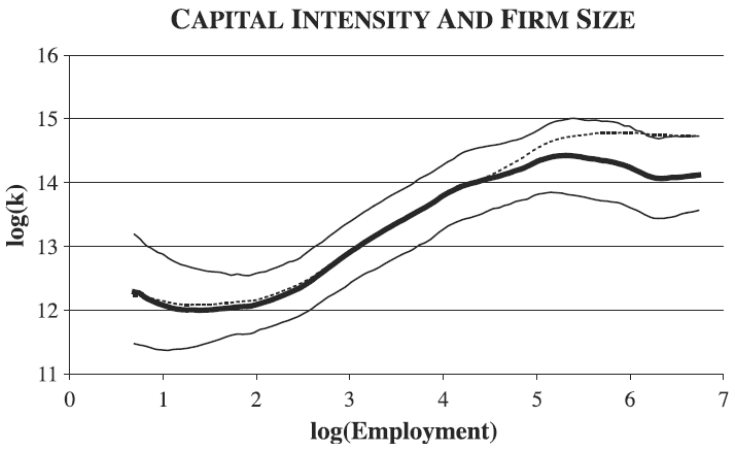


Fig. 2. Capital intensity and firm size. Note: The solid line shows the regression line of ln(k) on ln L. The kernel is Epanechnikov and the bandwidth is equal to 1.20. The thin lines indicate pointwise 95% confidence bands, calculated from 800 bootstrapped replications. To take the panel nature of the data into account, we bootstrapped from the firms rather than from the observations, which is a similar procedure to that used by Deaton (1997, pp. 216–218) for clustered data. The dashed line shows the regression line of ln(k) on ln L when k is not adjusted for worker quality heterogeneity as explained in the text.

(iii) Interpret the following figure (Figure 4 from Söderbom and Teal (2004)).

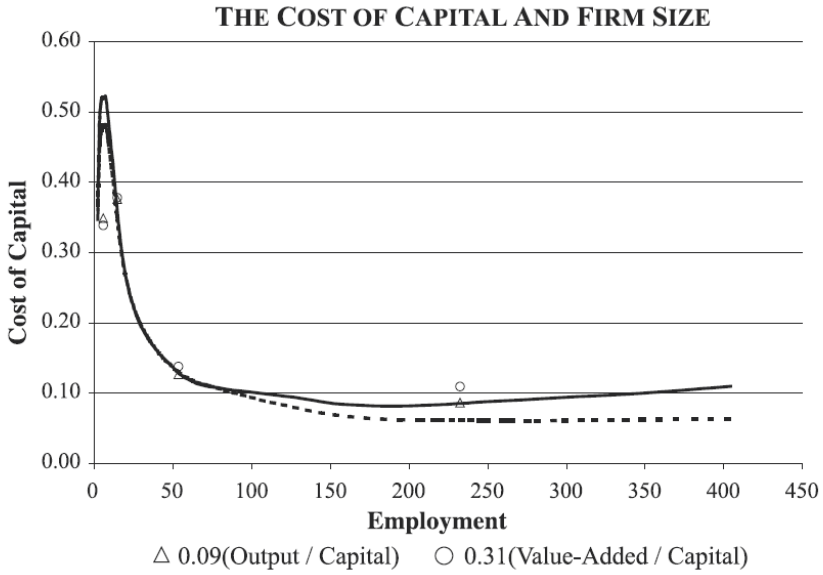


Fig. 4. The cost of capital and firm size. Note: The solid line shows the cost of capital based on the semiparametric regression of quality adjusted capital intensity on size. The dashed line shows the cost of capital based on the same regression without quality adjustment. The points indicated by △ and ○ indicate the marginal productivity of capital using the appropriate capital coefficient, evaluated at the mean values of log(Output/Capital) and log(Value-Added/Capital) reported in Table 1, by size category. The points are positioned horizontally at the mean values of employment for each size category.

(iv) The authors say this (p.390):

> Whether the implication of our findings is that substantial cost reductions and a more efficient use of capital are possible depends on the source of the differentials in factor prices. If factor prices vary due to differences in default risk, efficiency wage effects or other unavoidable consequences of information asymmetries and uncertainty then it may not be possible to reduce costs. If they reflect policy or labour market distortions, which are removable, then substantial cost reductions are possible.

Explain.