SEMI-PARAMETRIC ESTIMATION OF THE COMPANY GROWTH-SIZE RELATION

ESRC Centre for Business Research, University of Cambridge
Working Paper No. 32

Andy Cosh
ESRC Centre for Business Research,
Judge Institute for Management Studies, and
Queens' College
Silver Street
Cambridge CB3 9ET

Phone: 01223 335605
Fax: 01223 335566
E-Mail: adc1@eng.cam.ac.uk

Alan Hughes
ESRC Centre for Business Research
Department of Applied Economics
University of Cambridge
Sidgwick Avenue
Cambridge CB3 9DE

Phone: 01223 335248
Fax: 01223 335768
E-Mail: ah13@econ.cam.ac.uk

Hyehoon Lee
University of Leicester
Economics Department
University Road
Leicester LE1 7RH

Phone: 0116 252 5368
Fax: 0116 252 2908
E-Mail: h17@le.ac.uk

Stephen Pudney
University of Leicester
Economics Department
University Road
Leicester LE1 7RH

Phone: 0116 252 2887
Fax: 0116 252 5351
E-Mail: sep2@le.ac.uk

This Working Paper relates to the CBR Research Programme on Surveys, Databases and Database Management.

Abstract

We illustrate the potential of a number of recently-developed semi-parametric estimators, in the context of an analysis of the relationship between firm growth and mortality on the one hand and size on the other, conditioning the analysis also on age and industry dummies. These estimators produce results that suggest quite serious misspecification of the conventional firm mortality probit models, and slight but significant functional misspecification of the usual linear growth-size regression. The semi-parametric methods lead to a significantly different estimate of the joint mortality/growth distribution conditional on initial size age and industry.

Acknowledgements

We are grateful to the Economic and Social Research Council (grant H519255003) for financial support under its *Analysis of large and Complex Datasets* initiative, and under the *Surveys, Databases and Database Management* of the ESRC Centre for Business Research.

Key Words: Semi-parametric estimations; kernel estimation; company growth

SEMI-PARAMETRIC ESTIMATION OF THE COMPANY GROWTH-SIZE RELATION

1 Introduction

In the applied literature, a common finding of studies covering the last fifteen years or so has been that firm growth and failure rates decline with firm size and age (Schmalensee (1989), Hall (1987), Evans (1987a,b), Storey *et al* (1987), Dunne and Hughes (1994)). This is in contrast to studies covering earlier periods for the United Kingdom which showed that firm growth was positively related to size (Singh and Whittington (1968, 1975), Samuels (1965), Prais (1976), Hart (1965), Kumar (1984), Samuels and Chesher (1972)). Both sets of studies contradict a celebrated hypothesis attributed to Gibrat which holds that growth is independent of firm size. It has been argued that the result for earlier periods may reflect higher rates of growth by merger amongst larger companies (Hannah and Kay (1977), Hughes (1993)), whilst the result for more recent years may be attributable to selection bias.

When we estimate growth-size relationships with company panels, an unavoidable problem is sample attrition. Some companies may cease to exist during the period covered by the panel. Whereas slow growing large firms may simply slip slowly downwards through the size distribution for a considerable length of time before ceasing to trade, a smaller company is likely to hit the boundary of extinction much sooner. Small firms which have slow or negative growth may be more likely to disappear from the sample in any given time interval than are large firms. If slow growing small firms have a greater likelihood of failure than slow growing large firms, then estimates of growth by size, based on surviving firms only, will be biased towards finding a negative growth-size relationship. However, as we discuss below, there are some conceptual difficulties with this line of reasoning, and also serious identification problems. Recent applied studies have in any case found that a negative growth-size relationship remains even after correcting for possible sample selection bias (Dunne and Hughes (1994)).

The main focus of the present paper is methodological: to examine the

robustness of these results to decisions that are typically made about econometric specifications (usually linear regressions and probits) to be used in the analysis. We do this by implementing some recent semi-parametric estimators which are based on the Nadaraya-Watson kernel approach, and which generalise the standard regression and probit estimators. We focus on the relationship between two aspects of firms' development - their growth and failure rates - and firm size, allowing also for the influence of age and broad industrial sector. We do this for a large stratified sample of UK quoted and unquoted companies covering the whole coporate sector size range in the period 1976-1982. We find results which give grounds for concern about the use of standard linear parametric techniques.

The paper proceeds as follows. Section 2 discusses alternative econometric approaches and introduces a decomposition of the joint mortality/growth distribution which simplifies the application of semi-parametric estimators. Section 3 surveys semi-parametric estimation techniques available for estimation of each component of the joint distribution. Section 4 describes our data set. Sections 5 and 6 report and analyse estimation results produced by the parametric and semi-parametric approaches respectively. Section 7 compares the joint distribution of firm growth and mortality based on the alternative approaches and section 8 concludes.

2 Alternative approaches

Define the following notation: y_0 and y_1 are measures of the size of the firm in periods 0 and 1. If the firm dies between these two dates, y_1 is undefined, and a dummy variable ζ takes the value 1 if death occurs and 0 otherwise. A vector z contains a number of other variables to be used to explain or describe firm growth. Our objective is to study the distribution of $\{y_1 , \zeta\}$ conditional on $\{y_0 , z\}$, which underlies the evolution of the company population through time. Note that we are dealing here only with the processes of firm growth and death, not with the births of new firms.

Our first objective is to estimate the conditional probability density/mass function of $\{y_1 , \zeta\}$ conditional on $\{y_0 , z\}$. Since we are seeking to analyse the relation between growth and size (and death and size), we focus particular attention on the role of y_0. We are particularly anxious to avoid imposing *a priori* assumptions about the way in which y_0 enters the distribution. To estimate the distribution of $\{y_1, \zeta \mid y_0, z\}$ directly would be possible using classical parametric methods, notably maximum likelihood. These methods are widely available in standard software packages, and are consequently very widely used (see Dunne and Hughes (1994) for a typical example in the context of company growth). However, the parametric approach brings with it additional incidental assumptions (typically normality and linearity), which may have a critical influence on the results. Under the parametric approach, the joint distribution of $\{y_1 , \zeta\}$ would be specified as a partially-observed bivariate normal distribution with mean vector specified to be linear in y_0 and z. These assumptions could be relaxed within the parametric framework, for example by specifying flexible distributional forms (see, for example, Lee (1994), Gallant and Nychka (1987), Gabler, Laisney, and Lechner (1993)), and with more general specifications used for the conditional mean vector. However, these generalised specifications are not often used in practice.

An alternative approach is to use techniques that are non-parametric as far as possible. To introduce the idea of non-parametric regression, consider the following regression model, which ignores firm death and any other explanatory variables besides y_0:

$$\mathrm{E}(y_1 \mid y_0) \quad = \quad g(y_0) \qquad \textbf{(1)}$$

The non-parametric approach consists in estimating the entire function g(.) without restricting it to some parametric family such as the set of linear functions. The best-known of these methods use variants of the Nadaraya-Watson kernel smoothing technique (see Silverman (1986) and Härdle (1990) for surveys, and Deaton (1989) and Pudney (1993) for examples of their application in economics). The Nadaraya-Watson estimator of the value of the mean function $g(y_0)$ at an arbitrary point y_0 is as follows:

$$\hat{g}(y_0) = \frac{\sum_{j=1}^{n} k\left(\frac{(y_0 - y_{0j})}{\lambda_j}\right) y_{0j}}{\sum_{j=1}^{n} k\left(\frac{(y_0 - y_{0j})}{\lambda_j}\right)} \qquad (2)$$

where k(.) is a suitable kernel function (usually symmetric and non-negative), which integrates to unity. The estimator (2) can be interpreted as a sample analogue of the population regression $E(y_1 \mid y_0) = \int y_1 \, dF(y_0, y_1) / \int dF(y_0)$, where F is used as generic notation for a cumulative distribution function (cdf). It can also be viewed as a (rather sophisticated) smoothed form of a simple bar chart plotting group means of y_1 against ranges of values for y_0. The degree of smoothing in (2) is controlled by a set of bandwidth parameters λ_j which may vary (for instance, to give a higher degree of smoothing in regions where observations are sparse). Wherever we apply expressions like (2) in this study, we use the two-stage adaptive method of Breiman *et al* (1977) to generate bandwidths. It is convenient for our purposes to use a globally differentiable kernel such as the Gaussian pdf, to avoid difficulties with gradient-based optimisation techniques.

Estimators of this type have the enormous theoretical advantage of robustness against functional misspecification. However, there are circumstances in which kernel-type estimation performs poorly. One such case involves discrete explanatory variables. Although kernel regression has good asymptotic properties in such cases (see Bierens (1994)), it will nevertheless tend to work very badly in finite samples when the explanatory variables contain dummies which define a very fine partition of the sample into cells. For example, consider the following model where y_0 is continuous and (for simplicity) assume z is a vector of 0/1 dummies:

$$E(y_1 \mid y_0, z) = g(y_0, z) \qquad (3)$$

In this case, for a sufficiently small bandwidth, kernel estimation of g(.) is equivalent to computing separate non-parametric regressions of y_1 on

y_0 in each of the cells of the sample cross-classification defined by the dummies z. When there is a very large number of cells defined by the partition z, within-cell sample sizes will be small and thus the separate cell-specific regressions subject to high degrees of sampling error. This type of model is the rule rather than the exception in micro-econometrics, so this is an important drawback. A second limitation of kernel techniques is that their performance deteriorates very rapidly as we increase the number of explanatory variables. Again, high dimensionality tends to be a feature of micro-econometrics.

An intermediate approach is semi-parametric: we specify the statistical problem in such a way that it involves both non-parametric and parametric elements. In the previous regression example, this might amount to the following specification:

$$\mathrm{E}(y_1 \mid y_0, z) \quad = \quad g(y_0, z'\boldsymbol{\gamma}) \qquad (4)$$

leading to the problem of estimating the unknown bivariate function g(.) and unknown vector γ. This has two obvious advantages over (2): the dimension of g(.) is reduced and the linear form $z'\gamma$ is continuously-variable if z contains a continuous variable, and usually moves us much closer to continuous variability even if z is entirely discrete.

In practice there is no clear dividing line between the semi-parametric approach and a suitably general parametric approach. In the former, we treat the degree of flexibility of the estimated functional forms essentially as something to be estimated automatically from the data, while in the latter we treat it formally as fixed *a priori*, but in practice usually determine it by means of *ad hoc* model selection criteria (such as a score test for misspecification, or a likelihood-based criterion like that of Akaike). It is interesting to note that Gallant and Nychka (1987) proposed a technique based on a particular series-expansion as a semi-parametric method, with the order of the expansion treated as a quantity increasing with the sample size according to a pre-defined rule, whereas Gabler, Laisney, and Lechner (1993) implemented it as a fully-parametric technique with the order of the expansion fixed at a level suggested by model selection criteria. One could regard the latter

approach as a formal adaptive version of the former, in which case its statistical properties could in principle be derived. In the hands of an intelligent applied statistician, there may in any case be little to choose between the two interpretations.

The presence of firm death in the process complicates matters significantly, since it divides the sample into two separate regimes. Rather than estimate the distribution of $\{y_1, \zeta \mid y_0, z\}$ directly, we break it down into separate components which can be more easily estimated in a robust way using simple semi-parametric estimators. Consider the joint probability density/mass function[1] $f(y_0, y_1, \zeta \mid z)$, which has two components. Each component can be decomposed as follows:

Dying firms

$$f(\zeta = 1 \mid y_0, z) = \frac{f(y_0 \mid \zeta = 1, z)\, f(\zeta = 1 \mid z)}{f(y_0 \mid z)} \qquad (5)$$

Surviving firms

$$f(y_1, \zeta = 0 \mid y_0, z) = f(y_1 \mid y_0, \zeta = 0, z)\, \left(1 - f(\zeta = 1 \mid y_0, z)\right)$$

$$= f(y_1 \mid y_0, \zeta = 0, z) \left[\frac{f(y_0 \mid \zeta = 0, z)\, (1 - f(\zeta = 1 \mid z))}{f(y_0 \mid z)} \right] \qquad (6)$$

where we use the symbol f as generic notation for any pdf. The expressions (5) and (6) are particularly convenient for applied work, since the components on the right hand sides appear in forms that are relatively easy to estimate using semi-parametric methods.

Note that our objective here is only to estimate as flexibly as possible a conditional distribution. There is a further question, not addressed in this study, concerning the structural significance that can be attributed to such a distribution. It is well known, for example, that a dynamic model containing firm-specific random or fixed effects will be inconsistently

estimated by simple regressions, either on a full panel or a simple cross-section (see Nickell (1981), Arellano and Bond (1991), Pesaran and Smith (1995)). This is an issue concerning the estimation of distributions conditional not only on y_0, z, but also the unobservable effects. It does not affect our ability to estimate the purely observable distribution (5)-(6), which may be attributed a descriptive rather than structural role. Our aim here is not to construct a model that is structural in any sense, but to focus on other possible biases that might stem from the use of conventional linear regression and probit models, rather than more flexible forms.

3 Estimation techniques

There are four separate components of the distribution (5)-(6). The probability $f(\zeta=1 \mid z)$ is a binary response function of the sort which is conventionally estimated using maximum likelihood probit or logit analysis. The pdfs $f(y_0 \mid z)$ and $f(y_0 \mid \zeta=1, z)$ can be thought of as (not necessarily normal or linear) regressions of y_0 on z in the full population and the subpopulation of dying firms respectively. The conditional distribution $f(y_1 \mid y_0, \zeta=0, z)$ is again a regression relationship, this time defined on the population of surviving firms, but involving both y_0 and z as conditioning variables. We consider three separate estimation approaches for these three classes of relationship.

3.1 Binary response models

There is now a long list of available semi-parametric estimators for the simple binary response model, although relatively few applications of them as yet. Table 1 lists some of the semi-parametric estimators that have been proposed in the econometrics literature. There is a standard interpretation of the binary response model involving the specification of a latent indicator which generates a positive response by crossing a (zero) threshold. Thus:

$$\zeta = \pi(z'\boldsymbol{\beta} + \upsilon > 0) \tag{7}$$

where $z'\boldsymbol{\beta}$ is a fixed parameter vector (subject to an arbitrary normalisation); $\pi(A)$ is the indicator function, equal to 1 if the event A is true and 0 otherwise; and υ is an unobservable satisfying $E(\upsilon \mid z) = 0$ (or some analogous location restriction). Note that the distributional form of υ and the skedasticity function $\text{var}(\upsilon \mid z)$ are left unspecified.

An alternative model is the following direct specification for the response probability:

$$f(\zeta = 1 \mid z) = G(z'\boldsymbol{\beta}) \tag{8}$$

where $G(.)$ is an arbitrary non-negative function. This model is equivalent to the latent variable model only under special assumptions. For instance, when υ and z are independent and heteroskedasticity is ruled out, then $G(.)$ is the common cdf of $-\upsilon$ whatever value of z we choose to condition on. If there is heteroskedasticity in the latent variable model then the skedasticity function is $\text{var}(\upsilon \mid z) = \sigma^2(z)$ and:

$$f(\zeta = 1 \mid z) = G\left(\frac{z'\boldsymbol{\beta}}{\sigma(z)}\right) \tag{9}$$

where $G(.)$ is now the cdf of $-\upsilon/\sigma(z)$. Thus (8) is not generally valid as a representation of (7) unless σ happens to be a function of the same linear form $z'\boldsymbol{\beta}$ as appears in the latent indicator. Even in that case, the resulting function of $z'\boldsymbol{\beta}$ that gives the response probability will not necessarily have the properties of a cdf, as is assumed in probit or logit analysis. Note that it is not always possible to express the model (8) in the form of a heteroskedastic latent indicator model (9), since the implied function $\sigma(z)$ may not exist or may be negative.

The estimators detailed in Table 1 are all based on the optimisation of some objective function. Specifically:

MSCORE

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(2\,\pi(\zeta_i = 1) - 1 \right) \pi\left(z_i'\boldsymbol{\beta} \geq 0 \right) \qquad (10)$$

Smoothed MSCORE

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(2\,\pi(\zeta_i = 1) - 1 \right) \Phi\left(\frac{z_i'\boldsymbol{\beta}}{\mu_n} \right) \qquad (11)$$

Semi-parametric ML

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left(\max_{G(.)} \sum_{i=1}^{n} \left(\zeta_i \ln G(z_i'\boldsymbol{\beta}) + (1-\zeta_i) \ln\left(1 - G(z_i'\boldsymbol{\beta}) \right) \right) \right) \qquad (12)$$

Klein-Spady

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(\zeta_i \ln \hat{G}(z_i'\boldsymbol{\beta}) + (1-\zeta_i) \ln\left(1 - \hat{G}(z_i'\boldsymbol{\beta}) \right) \right) \qquad (13)$$

Maximum rank correlation

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j<i} \left(\pi(\zeta_i > \zeta_j)\ \pi(z_i'\boldsymbol{\beta} > z_j'\boldsymbol{\beta}) + \pi(\zeta_i < \zeta_j)\ \pi(z_i'\boldsymbol{\beta} < z_j'\boldsymbol{\beta}) \right) \qquad (14)$$

In these expressions, μ_n is a bandwidth sequence, $\Phi(\,.\,)$ is the normal cdf, and $\hat{G}(.)$ is a non-parametric kernel regression of ζ on $z'\beta$.

Table 1 Estimation methods for the binary response model

Technique	Reference	Comments
Maximum score estimator (MSCORE)	Manski (1975)	Estimator of scaled coefficients of latent regression only; no assumption of homoskedasticity; unknown limiting distribution; convergence at rate slower than $\sqrt{n}$
Smoothed MSCORE	Horowitz (1992)	Estimator of scaled coefficients of latent regression only; no assumption of homoskedasticity; convergence at rate $\sqrt{n}$ to normal limiting distribution; asymptotically efficient in the absence of further assumptions
Semi-parametric maximum likelihood	Cosslett (1983)	Estimator of α and $G(\,.\,)$; assumes homoskedasticity ($\sigma(z) = \sigma$); convergence at rate $\sqrt{n}$ to normal limiting distribution; asymptotically efficient under homoskedasticity.
Empirical likelihood	Klein and Spady (1993)	Estimator of α and $G(\,.\,)$; assumes homoskedasticity; convergence at rate $\sqrt{n}$ to normal limiting distribution; asymptotically efficient under homoskedasticity.
Maximum rank correlation	Han (1987)	Estimator of α only; no assumption of homoskedasticity; convergence at rate $\sqrt{n}$ to normal limiting distribution.

3.2 Regressions on z

The most general form of semi-parametric regression would involve the following conditional expectation function:

$$\mathrm{E}(y_0 \mid z) = g(z'\beta) \qquad (15)$$

where g(.) is some unknown function and $\boldsymbol{\beta}$ is again a normalised parameter vector.[2] Models of this kind are not easy to estimate. If there is a continuously-distributed variable in the z vector, then an appealing approach is to use the following kernel least squares estimator. For any given value $\boldsymbol{\beta}$, we compute the non-parametric kernel regression of y_0 on the constructed index variable $z'\boldsymbol{\beta}$ and evaluate the residual sum of squares. An optimisation algorithm is then used to minimise the residual sum of squares with respect to $\boldsymbol{\beta}$. Thus:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_{0i} - \hat{g}(z_i'\boldsymbol{\beta}) \right)^2 \qquad (16)$$

where $\hat{g}(.)$ is the non-parametric regression of y_0 on $z'\boldsymbol{\beta}$.

3.3 Regressions on y_0 and z

Ideally, one would like to estimate the distribution $f(y_1 \mid y_0, \zeta=0, z)$ by fitting a very flexible model of the following form:

$$\mathrm{E}(y_1 \mid y_0, \zeta=0, z) = g(y_0, z'\boldsymbol{\beta}) \qquad (17)$$

where $g(.,.)$ is an unknown function and $\boldsymbol{\beta}$ is a (normalised) parameter vector. This specification suggests a least-squares estimator that minimises the following criterion:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_{1i} - \hat{g}(y_0, z_i'\boldsymbol{\beta}) \right)^2 \qquad (18)$$

In the application discussed below, we simplify this procedure still further by introducing an additivity assumption so that the model becomes:

$$E(y_1 \mid y_0, z) = g(y_0) + z'\beta \tag{19}$$

The reason for this is that trivariate kernel regression produces estimates that converge at a slower rate than bivariate kernel regression ($O_p(n^{-1/3})$ rather than $O_p(n^{-2/5})$), so that if the additivity assumption is valid the properties of the estimator should be considerably improved.

Least squares estimation of this type of model is equivalent to the estimator discussed by Robinson (1988) (see also Härdle (1990, section 9.1)). The additive semi-parametric regression estimator can be motivated by an analogue of the formulae for partitioned linear regression, where:

$$\hat{\beta} = \left(\hat{Z}'\hat{Z}\right)^{-1} \hat{Z}'\hat{y}_1 \tag{20}$$

$$\hat{g} = P(y_1 - Z\hat{\beta}) \tag{21}$$

where Z, X and y are data matrices, $\hat{y} = y - Py$ and $\hat{Z} = Z - PZ$ are the residuals from regressions of Z and y_1 on y_0 and P is the projection matrix $y_0(y_0'y_0)^{-1}y_0'$. In the semi-parametric case, the transformation P is replaced by a kernel smoother, P^*, with i,j th element:

$$p_{ij}^* = \frac{k\left(\dfrac{(z_i - z_j)'\beta}{\lambda_j}\right)}{\sum_{r=1}^{n} k\left(\dfrac{(z_i - z_r)'\beta}{\lambda_r}\right)} \tag{22}$$

Experimentation has revealed little need in this case for the use of a trimming function as suggested by Robinson (1988) to eliminate the influence of observations with small estimated x-densities.

4 Data

Our estimates make use of a dataset prepared at the ESRC Centre for Business Research, Cambridge. The dataset is based on a size stratified sample of U.K. companies whose computerized accounts for the years 1976 - 1982, prepared by the Business Statistics Office, Newport were available via the ESRC Data Archive at the University of Essex. We describe the salient features of this sample and the way in which we have checked and augmented it for the analysis reported in this paper (and for the analysis in companion papers dealing with the economics of company failure and acquisition activity (Cosh and Hughes (1994, 1995), Cosh, Hughes and Kambhampati(1993)). This has involved, *inter alia*, the analysis of the microfiche records of several hundreds of individual company records obtained from Companies House, London and Cardiff to trace the fate of companies leaving the sample in the course of the sample period; to augment the accounting records of sample companies with details of company age; to obtain information on the characteristics of directors; and to extract accounting data for acquired companies in the aftermath of acquisition. The first two of these data augmentation exercises are directly relevant to this paper.

4.1 The Business Statistics Office Sample

The sample of companies whose accounts form the basis of our analysis is one of a number of samples which have been constructed by the Board of Trade and successor government departments in the aftermath of the 1948 Companies Act to assist in the National Accounts analysis of company profitability and finance. Summary statistics relating to these samples have appeared regularly in *Business Monitor MA3 -Company Finance* and its predecessor publications. Until the late 1970s the analysis was based on a panel of the largest (primarily quoted) companies which was periodically rebased to exclude companies falling below certain size criteria and to include companies newly rising above them (Lewis (1979)). This exclusion of smaller companies was then rectified by the construction of a size-stratified sample of around 3000 independent industrial and commercial companies drawn from the official list of all companies registered in Great Britain in 1975. This sample, which we have used as the basis for our work,

consisted of the largest 500 companies, a sample of one in two of the next 1000 largest, a sample of one in seventy of "medium" sized companies, and a one in three hundred and sixty sample of the remainder. This sample was in turn replaced, in the mid 1980s, by a futher sample of around 3000 companies, again based on the Companies House Register, but including the top 2000 companies in 1981 and a one in three hundred sample of the remainder. After minor adjustment to the larger companies component of this sample in 1985 there was a major rebasing in 1990 to ensure the inclusion of the largest 2000 companies of 1987.

Of these three potential sources of company data the first contains very few small and medium sized companies and covers a relatively distant historical time period. Of the other two the third and most recent sample appears, at first, to offer a number of potential advantages. First, compared to both of the other samples, it is the most comprehensive in its treatment of births and deaths. This is partly because of improvements in the management of the underlying records system at Companies House. It has been estimated that in the face of chronic failure to submit accounts on time and pressures arising from keeping pace with the high rates of business registration of the late seventies and early eighties the Companies Registers contained accounts for less than half of all companies which should have been filing by 1984/85 (Knight (1987)). On the other hand the increasing use by small and medium sized companies of dispensation to submit modified accounts has greatly reduced the range of data available for small and medium sized companies. The analysis in this paper focuses therefore on the stratified sample drawn in 1975.

For each of these sampled companies our accounting data is available in principle from 1976 to 1982 or until their exclusion from the panel before 1982. In addition to accounting data each company in the panel has a number of indicators, including indicators of "death" either by acquisition, liquidation or other cause. For the purpose of this paper we excluded 168 property companies because of accounting inconsistencies with other industrial and commercial companies in term of asset valuation. Companies with total assets of less than £50,000 were also excluded from our analysis of growth because of the high incidence of missing data and inconsistent asset records in this size class. The final sample for growth analysis consists of 2142

companies of which 527 'died' between 1976 and 1982 and 1615 were 'alive' in both these years.

In the analysis which follows we use total assets as our measure of size (S); this is because sales figures are not universally available and the relatively high ratio of current liabilities to total assets amongst small companies made net assets an inappropriate measure of size for a small but significant portion of our sample.

Figure 1 shows a non-parametric kernel estimate of the initial size distribution (in log form). This reveals a very strongly bimodal distribution, reflecting the non-uniform sampling scheme used originally by the Board of Trade. The strong rightwards shift in the distribution between 1976 and 1982 is the result of both company growth in real terms and inflation, since assets are valued in nominal historic cost terms. Figure 2 shows a similar non-parametric estimate of the distribution of log age in 1976. Possibly as a result of the non-uniform sampling, there are multiple modes at 6, 16 and 43 years.

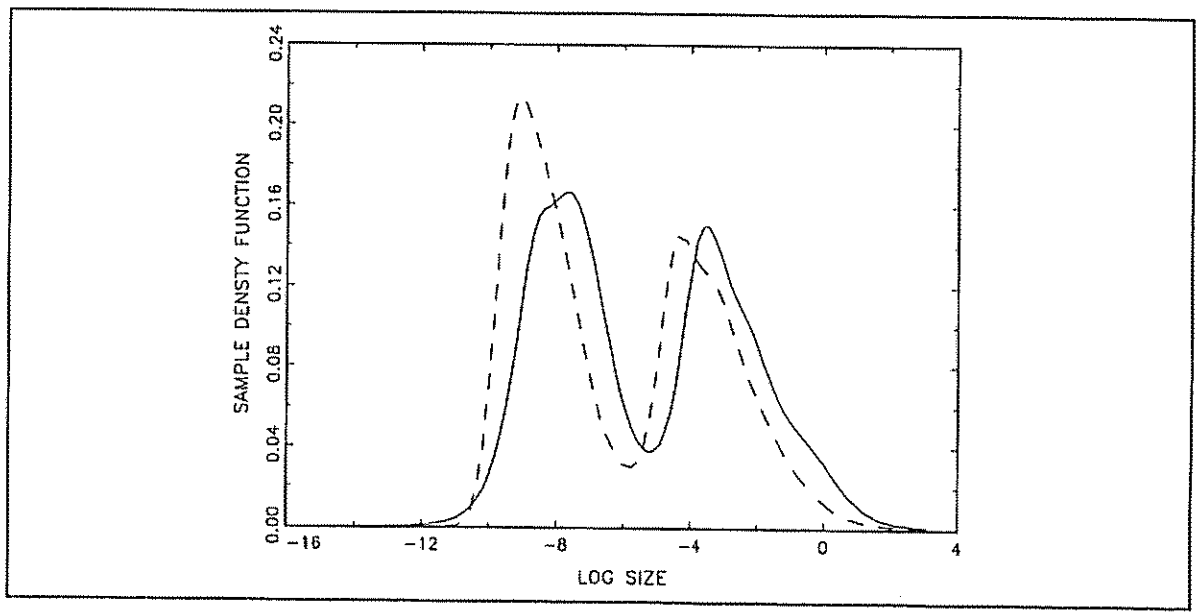


Figure 1 Distributions of 1976 and 1982 firm size (ln £ billion asset value; _ _ _ = 1976; ____ = 1982)

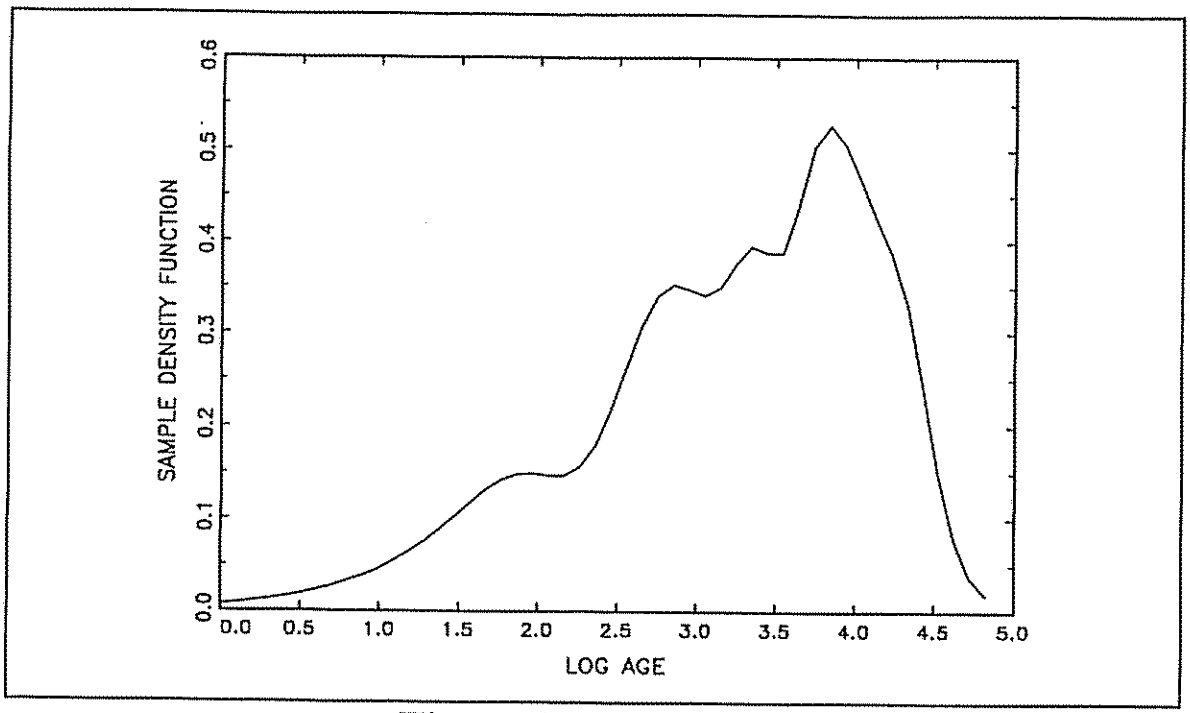


Figure 2 Distribution of firm age

Table 2 provides a summary of the dynamics of growth and company "death" in our sample, in the form of a matrix cross-classifying companies by opening and closing total asset size class. Inflation creates a tendency for cell frequencies to be higher to the right of the main diagonal. The matrix nevertheless reveals a clear clustering of firms along the diagonal. Most firms therefore remain in their opening size class (in the case of largest size class only demotions are of course possible). There are a handful of cases of extremely rapid growth with five companies, for instance, whose 1976 total assets fell in the £100,000-£500,000 range, increasing in size to over £2.5m by 1982. The matrix also reveals a tendency for death rates to fall once the £2.5m size boundary is crossed. For the largest size class the failure rate is approximately one half of that in the 3 smallest size classes. A separate analysis (Cosh and Hughes (1994)) shows that the higher death rate of the smallest size classes is primarily due to liquidation and bankruptcy, since merger deaths are relatively less significant for them than for the middle sized companies.

Table 2 The distribution of sample companies cross classified by opening and closing size, survival and death

1976 SIZE CLASS	1982 SIZE CLASS										
	Total	<50	50 -100	100 -500	500- 2500	2500 -12500	12500 -62500	>62500	Total Survivors	Total Deaths	Death Rate %
50-100	305	21	58	132	7	0	0	0	218	87	28.5
100-500	16	3	10	249	182	4	0	0	448	168	27.3
500-2500	235	0	0	5	124	36	2	0	167	68	28.9
25000-12500	256	0	0	0	4	59	124	2	189	67	26.2
12500-62500	431	0	0	0	2	6	214	114	336	95	22.0
62500-	299	0	0	0	1	0	9	247	257	42	14.1
TOTAL	2142	24	68	386	320	105	349	363	1615	527	24.6

* Size is measured in £'000.

Table 3 Mean growth rates by size of company: survivors only

	Whole Sample	50-100	100-500	500-2500	2500-12500	12500-62500	>62500
No. of Survivors	1615	218	448	167	189	336	257
	percentage points over 1976-1982						
Mean	121	139	142	130	114	99	100
Median	86	76	96	84	91	80	81
St. Dev.	170	233	189	231	131	103	107
Kurtosis	4715	1722	2688	7081	1014	402	535

* Size groups measured by total assets in 1976 (size £'000).

Table 4 Significance Tests of Differences in Mean and Median Growth Rates

Size (£'000)	100-500	500-2500	2500-12500	12500-62500	over 62500
50 - 100	- --	+ -	+ -	+++ -	+++ -
100 - 500		+ +	+++ +	+++ +	+++ -
500 - 750			+ -	++ +	+ +
2500 - 12500				+ +	+ +
12500 - 62500					- -

Notes: + row av. > column av.; ++ sig. at the 10% level; +++ sig. at the 5% level
- row av. < column av.; --sig. at the 10% level; --- sig. at the 5 % level
Upper entry in each cell is the test of means and the lower entry is the test medians

The pattern of variations in growth rates of the surviving companies by total asset size class in 1976 is shown in Table 3. In addition to the mean and median of growth rates for all survivors and for each of our 6 size classes this table also reports measures of dispersion, skewness and kurtosis. Table 4 reports a matrix summarizing tests of significance of differences in mean and median survivor growth rates across all pairs of size classes. This analysis reveals that the smallest 3 size classes have much greater variance of growth rates than the 3 largest. It also shows that the distribution of growth rates is

positively skewed especially in the smaller size classes, and that the median is therefore a more appropriate measure of central density than the mean. The analysis of differences in means across size classes suggests that the smallest two size classes have significantly higher growth rates than the two largest size classes. In addition it shows that size class 2 has a higher mean growth rate than size class 4, and that a similar though less significant result applies in a comparison of size class 4 against size class 5. However, none of these results is significant when we compare median growth rates. In fact, only one significant difference emerges, and that is the superior median growth rate at the lower end of size distribution where size 2 compares favourably with size class 1.

5 Parametric Estimation Results

5.1 Linear regression and probit

Conventional estimation of relationships like those embodied in the distribution defined by equations (5) and (6) above is based on linear regression and probit analysis. The first four columns of Table 5 give the results of applying these methods. The first column gives the regression of y_1 on y_0 and z, using the sample of surviving firms. The coefficient of y_0 is not significantly different from unity, implying that, in this conditional sense, Gibrat's law cannot be rejected. Growth is negatively related to the age of the firm, and there are significant industry effects, with industries 2 and 4 (miscellaneous manufacturing and services) tending to grow faster over the period. The degree of fit is extremely high for a cross-section relation, with an R^2 of 0.95.

A conventional approach to testing the validity of a cross-section regression model like this is to use Ramsey's RESET test (Ramsey (1969)), which involves using the square of the fitted value as an additional regressor and testing its relevance using a t-test (in this case used in its asymptotic χ^2 form). The low value of the RESET test statistic indicates no evidence of misspecification of functional form. Heteroskedasticity is part of the maintained hypothesis underlying the RESET test, and we investigate this using a score test of the null hypothesis of homoskedasticity, which involves essentially a regression of the squared residual on the square of the fitted

value with an intercept. The score test then amounts to a t-test on the slope coefficient (again presented in asymptotic χ^2 form here). Once more there is no significant evidence of heteroskedasticity, so this conditional growth regression appears to conform to the assumptions of classical linear regression in these two important respects.

Columns 2 and 3 of Table 5 give the regressions of initial size, y_0, on log age and the industry dummies, for the full sample and the subset of dying firms respectively. Age has a strong positive effect, with an elasticity of 1.0 or greater. Relative to the norm, industries 1 and 2 (engineering and manufacturing) tend significantly towards large firm sizes and industry 3 (retailing etc.) tends towards small firms. The degree of fit is low, with R^2 below 0.3 in both cases: a result that simply confirms the heterogeneity of the stock of firms in existence at a given time. For these two regressions, conventional specification tests give a more negative verdict. The RESET test is highly significant, indicating that the distribution of initial firm size (whether conditioned on subsequent survival or not) has a nonlinear conditional mean function. There is no indication of heteroskedasticity, but the result of the test may be distorted by misspecification of the mean.

The fourth column of Table 5 gives the results of fitting a probit model for the probability that the firm dies between 1976 and 1982. This probit has log age and the industry dummies as explanatory variables; firm size is excluded, so that the result is directly comparable with the semi-parametric estimates of the distribution $f(\zeta=1 \mid z)$ appearing in equation (5). We find that the probability of death declines with the age of the firm, but that there are no significant industry effects. We have computed an analogue of the RESET test for this model by testing the significance of the constructed variable $(z'\hat{\beta})^2$ as an additional explanatory variable. Again, we apply the RESET specification, in the form of a score test for the irrelevance of an additional explanatory variable defined as $(z'\beta)^2$, and find a rather marginal result that causes some concern. The null hypothesis is rejected at a 10% significance level, but not at the more conventional 5% level.

Table 5 Parametric Estimation Results (Standard errors in parentheses)

Covariate	Linear regression / probit				Selectivity model	
	y_1 [a]	y_0 [b]	y_0 [c]	Pr(death)[b]	y_1 [a]	Pr(death)[b]
Constant	0.772 (0.067)	3.139 (0.210)	3.810 (0.366)	-0.365 (0.115)	0.766 (2.154)	-0.216 (0.122)
y_0	0.991 (0.007)	-	-	-	0.991 (0.067)	-0.045 (0.012)
ln(age)	-0.052 (0.020)	1.310 (0.063)	1.040 (0.112)	-0.122 (0.032)	-0.052 (0.111)	-0.067 (0.036)
Industry 1	0.035 (0.051)	0.873 (0.173)	0.658 (0.325)	0.069 (0.095)	0.035 (0.170)	0.107 (0.096)
Industry 2	0.113 (0.048)	1.043 (0.161)	0.737 (0.313)	0.089 (0.089)	0.112 (0.206)	0.136 (0.091)
Industry 3	0.054 (0.051)	-0.306 (0.168)	-0.232 (0.317)	0.110 (0.095)	0.055 (0.159)	0.097 (0.095)
Industry 4	0.185 (0.059)	-0.079 (0.199)	-0.545 (0.350)	0.007 (0.115)	0.184 (0.061)	0.0006 (0.111)
Number of observations	1615	2142	527	2142	2142	2142
σ	0.642	2.428	2.292	-	0.641 (0.023)	-
RESET test $\chi^2(1)$	0.067	115.88	22.468	2.792	-	3.580
LM heteroskedasticity $\chi^2(1)$	0.227	0.696	0.645	-	-	-
Selectivity correlation	-	-	-	-	0.014 (5.224)	-
R^2	0.95	0.260	0.214	-	-	-

[a] based on the sample of surviving firms only;
[b] based on the full sample;
[c] based on the sample of non-surviving firms.

Assuming homoskedasticity, we use these four parametric estimates in section 7 below to construct an estimate of the distribution (5)-(6), as an alternative to the semi-parametric estimates presented in section 6. Before turning to these, we consider the possible use of estimates "corrected" for sample attrition.

5.2 Adjustments for attrition

The regression of y_1 on y_0 and z presented in the first column of Table 5 is conditioned on the event that the firm survives through the period 1976-82. There is a large literature in econometrics dealing with the problems of estimating a relationship (such as the growth-size relation) when there exists another correlated process (in this case company death), which randomly excludes individuals from the sampling procedure. Following the work of Heckman (1974) and Hausman and Wise (1979), ML and 2-step parametric estimators are widely available for the following model:

$$y_1 = \beta_1 y_0 + z'\beta_2 + u \qquad (23)$$

$$y_1 \text{ observed if } \gamma_0 y_0 + z'\gamma + v > 0 \qquad (24)$$

where u and v have a correlated bivariate normal distribution. Note that equation (24) specifies a linear probit model that is a direct specification of the conditional probability defined by equation (5) above. Thus the final column of Table 5 can be regarded as a conventional estimate of (5).

This selectivity model is widely used, and selectivity-corrected estimates of the firm growth-size relation have been published by Dunne and Hughes (1994). Although firm death can indeed be interpreted as a selection process, it is not obvious that the model (23)-(24) makes sense in this context. In plain language, (23) says that there is a simple relationship that governs the tendency of firms to grow through time, while the probit (24) states that there is another separate relationship determining whether or not a firm dies. An estimate of (23) thus amounts to a prediction of the rate of growth that a firm either actually achieves or (if it in fact dies) would have achieved. One might argue that this kind of prediction is very hard to interpret in any useful way. In some (but not all) cases[3], the processes of growth and death are inseparable. Unsuccessful businesses may die because they encounter difficult periods when demand (and thus size) contracts. On this view, negative growth and death are simply two aspects of the same event, and it

makes little sense to ask what would have happened to firm growth if the possibility of death had somehow been artificially removed.

Whether or not these reservations are accepted, the ML selectivity estimates appearing in the last two columns of Table 5 add little to the simple regression in the first column (see Dunne and Hughes (1994) for a similar result). The estimated correlation between u and v turns out to be very small, with a huge standard error, and consequently the selectivity-corrected firm size regression is virtually identical to the uncorrected regression. Essentially, the selectivity correlation is unidentifiable in this case (as in many other practical applications), and we pursue the approach no further.

6 Semi-parametric Estimation Results

As discussed in section 3, we estimate four separate components of the distribution (5)-(6); the binary response function for ζ; regressions of y_0 on z for the whole sample and for dying firms; and a regression of y_1 on y_0 and z.

6.1 The model for probability of death

There are still rather few published applications of semi-parametric estimators for the binary response model. In modelling the probability of a firm's death during the sample period, we initially considered three approaches: Horowitz' (1992) smoothed maximum score estimator; Ichimura's (1993) semi-parametric least squares estimator applied to a regression of ζ on z; and Klein and Spady's (1993) empirical maximum likelihood estimator. We encountered serious problems in optimising the objective function for Horowitz' estimator, because of the existence of numerous local optima. Even the use of very good starting values in Monte Carlo experiments failed to achieve reliable convergence, and there is a clear need for robust global optimisation algorithms (such as the method of simulated annealing used by Horowitz (1993)) in computing this estimator. We finally opted for the Klein-Spady estimator, which is asymptotically efficient under fairly restrictive assumptions. However, Monte Carlo experiments suggest that Ichimura's (inefficient) least-squares estimator performs nearly as well in finite samples.

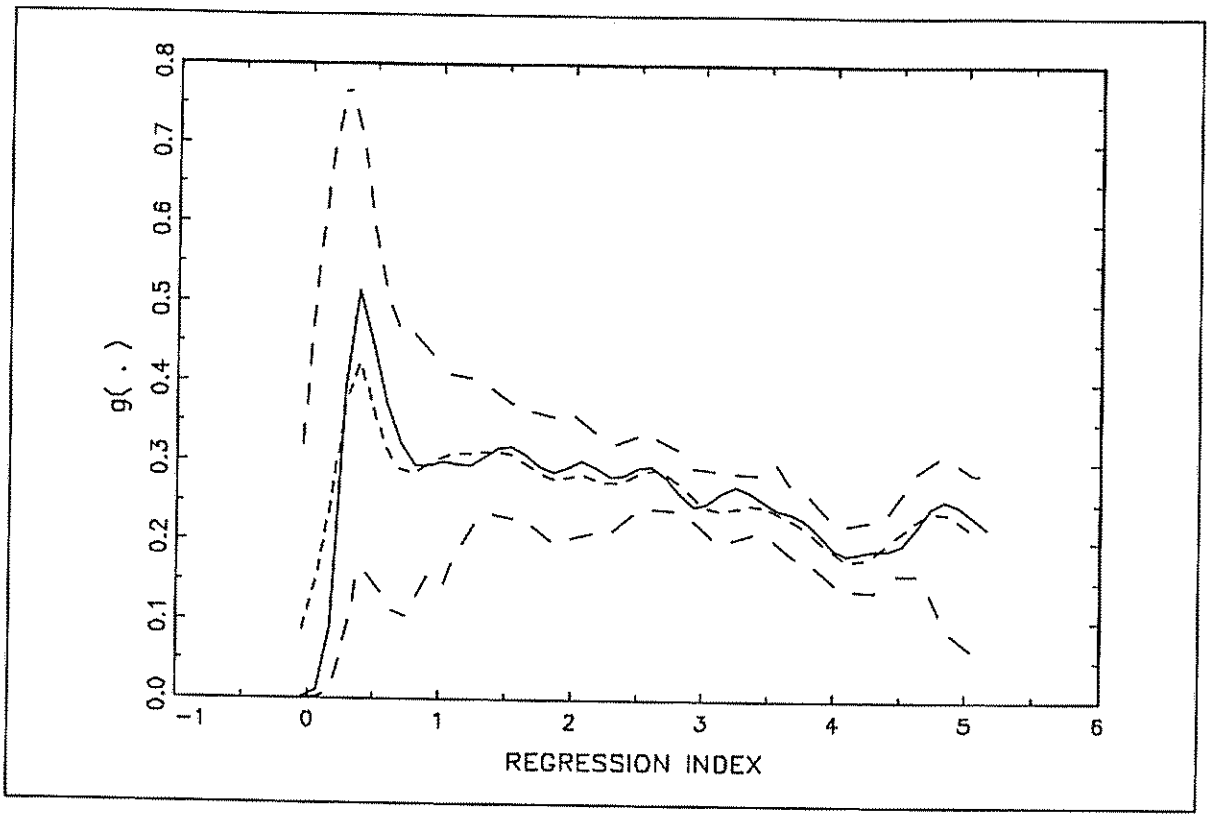


Figure 3 Klein-Spady Estimation Results
(____ = result from full sample; - - = 90% empirical confidence interval; - - - = mean of bootstrap replications)

Table 6 Klein-Spady estimation results
(Standard errors from 40 bootstrap replications)

Covariate	Coefficient	Standard error	Normalised linear probit*
Log(age)	1.000	-	1.000
Industry 1	-0.005	0.230	-0.566
Industry 2	0.566	0.219	-0.730
Industry 3	0.074	0.156	-0.092
Industry 4	-0.004	0.154	-0.057

* Coefficients from table 5; each coefficient divided by that of log(age).

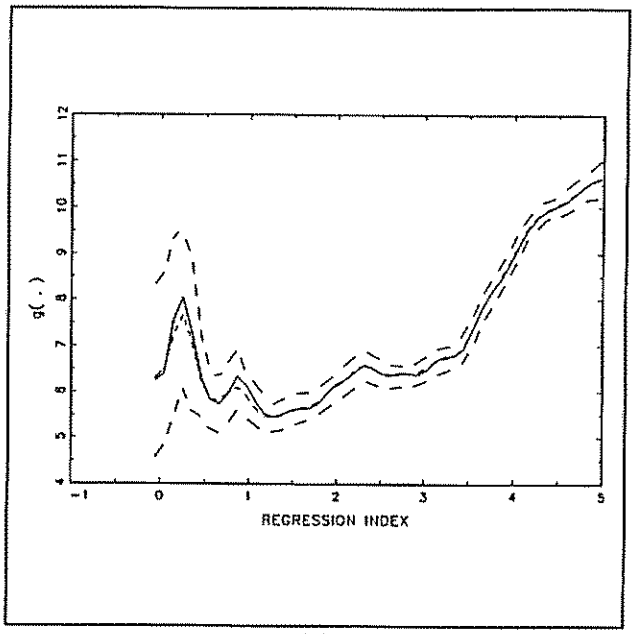
Figure 3 and Table 6 report the results of applying Klein and Spady's estimator. When we normalise the coefficient of log age to be unity, the estimate of the unknown function $G(.)$ in (8) turns out to be a decreasing function of regression index $z'\beta$ for most of the range. Thus, as in parametric estimation, the probability of firm death is predicted to decline with age. For comparison, we also include in Table 6 the coefficients from an analogous probit model, rescaled so that the coefficient of log age is unity. In the semi-parametric case, the second of the industry dummies produces a significant effect, implying a lower inherent survival rate for firms in the non-engineering sectors of manufacturing industry. In contrast, the corresponding linear probit estimate is not statistically significant.

6.2 Regressions of y_0 on z

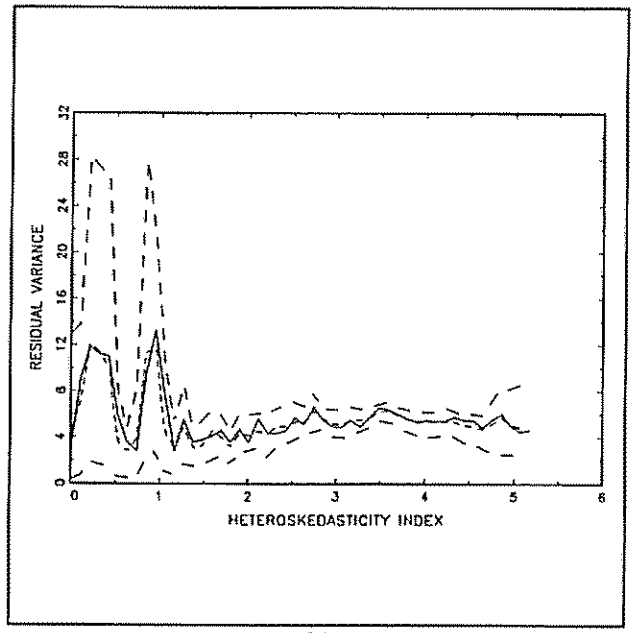
Table 7 Semi-parametric regression results for full sample (Standard errors computed from 40 bootstrap replications)

Covariate	Regression (dependent variable = log initial size)		Skedasticity (dependent variable = squared residual)	
	Coefficient	Standard Errors	Coefficients	Standard Errors
Log(age)	1.000	-	1.000	-
Industry 1	0.239	0.084	0.338	0.206
Industry 2	0.359	0.064	0.194	0.144
Industry 3	-0.114	0.061	0.635	0.374
Industry 4	-0.060	0.079	0.661	0.374
R^2	0.318		0.021	

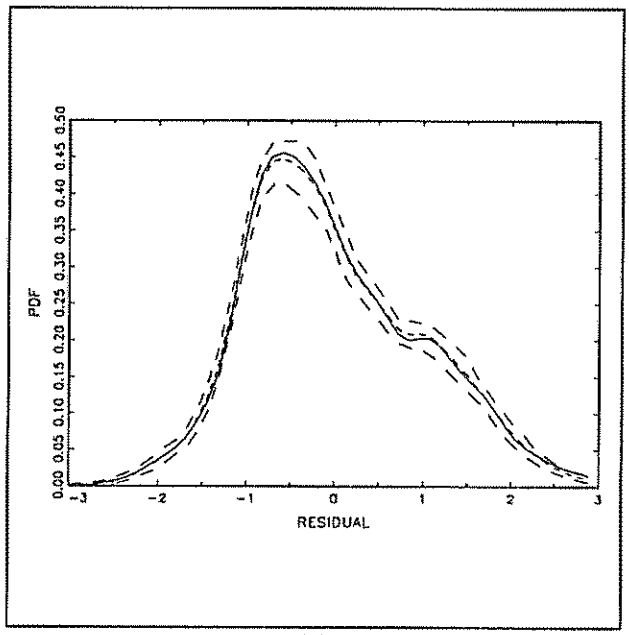
We estimate the regression of y_0 on z, $g(z'\beta)$, using Ichimura's (1993) semi-parametric least squares estimator first for the whole sample. The results are presented in Figure 4(a) and Table 7. The estimates of the unknown function g(.) turns out to be nearly convex and increasing over most of the range of



(a)



(b)



(c)

Figure 4 Full-sample semi-parametric regression for initial firm size
(____ = regression with full sample;
_ _ _ = 90% confidence interval;
- - - = bootstrap mean)

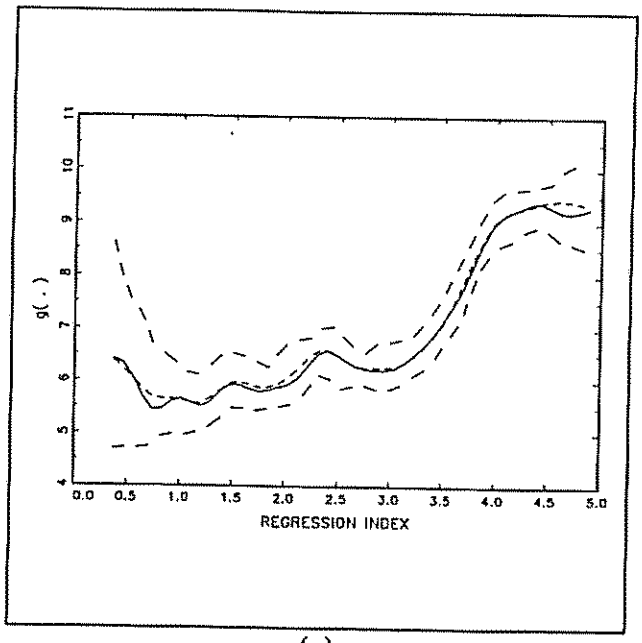
$z'\beta$. The 90% bootstrap confidence interval suggests that g(.) is estimated with good precision and is significantly nonlinear. The implied age-initial size relation (after controlling for industry effects) appears predominantly positive, as in the linear regression of Table 5. There are three significant industry effects, with industries 1 (metal/engineering) and 2 (miscellaneous manufacturing) characterised by larger firm sizes and industry 3 (retailing etc.) by smaller firm sizes. Again, this conclusion is qualitatively similar to that of the linear regression, but the relative effects of age and industry are quite different. In Table 5, the ratios of the coefficients of industry dummies to that of log age are more than double the comparable semi-parametric coefficients in Table 7. The fit as measured by R^2 is also considerably improved. Thus semi-parametric estimation does appear to offer a real improvement in flexibility over linear regression here.

In addition to the regression function itself, we have also estimated a skedasticity function by computing a similar semiparametric regression, using the squared residual as a dependent variable. This procedure estimates the following function:

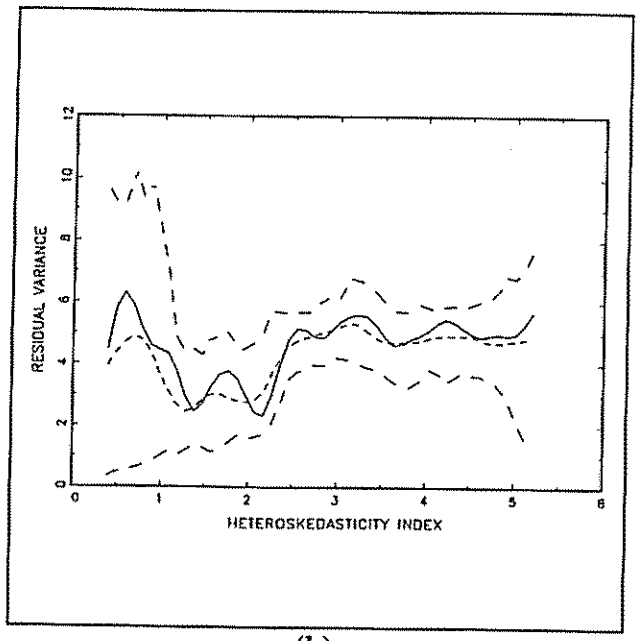
$$var(y_0 \mid z) = \psi(z'\gamma) \qquad (25)$$

The estimate of $\psi(.)$ is plotted in figure 4(b) and the bootstrap confidence intervals suggest that there is no significant variation in the residual variance. The low R^2 and insignificant coefficients in the estimate of γ also confirm this. Finally, figure 4(c) plots a kernel estimate of the density function of the regression residual. There is rather strong evidence of positive skewness in the distribution.

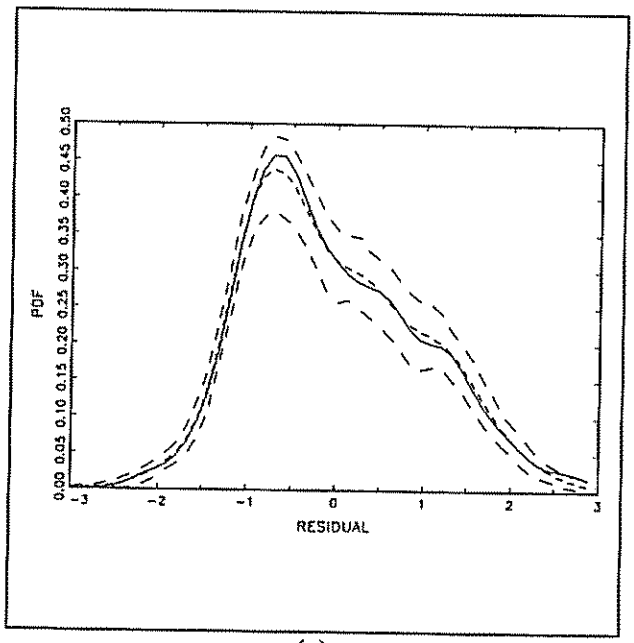
Figure 5 and Table 8 show the results of an analogous semi-parametric regression for initial size in the subsample of firms which are observed to die before the end of the observation period. The results are similar, with strong evidence for an increasing and mildly convex relationship. Industry effects are less clear, with only industry 2 (miscellaneous manufacturing) implying a strongly significant increase in size relative to other industries. Again, the semi-parametric regression yields a much better fit than the linear regression in Table 5, and also estimated industry effects that are much smaller relative to the age effect.



(a)



(b)



(c)

Figure 5 Dying firms: semi-parametric regression for initial size
(____ = sample regression;
_ _ _ = 90% confidence interval;
- - - = bootstrap mean)

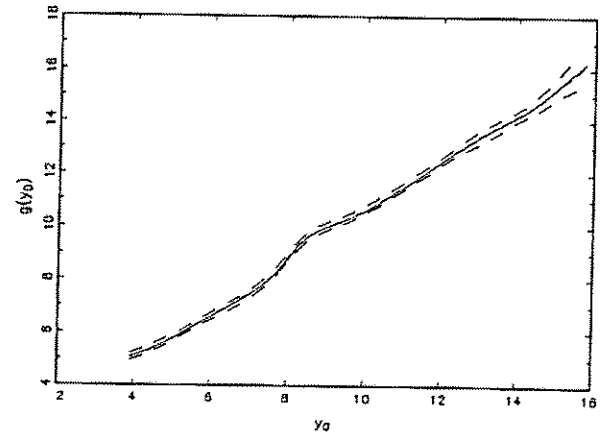
Table 8 Semi-parametric regression results for subsample of dying firms (Standard errors computed from 40 bootstrap replications)

Covariate	Regression (dependent variable = log initial size)		Skedasticity (dependent variable = squared residual)	
	Coefficient	Standard error	Coefficient	Standard error
Log(age)	1.000	-	1.000	-
Industry 1	0.105	0.103	-0.303	0.199
Industry 2	0.365	0.132	0.700	0.263
Industry 3	-0.140	0.116	0.485	0.231
Industry 4	-0.327	0.182	-0.286	0.177
R^2	0.276		0.034	

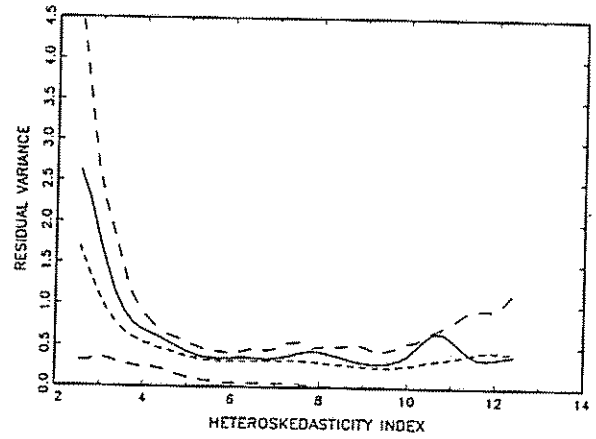
The evidence on heteroskedasticity is quite mixed in this case. There is some evidence for a lower residual variance for firms near the bottom of the estimated range of values of $z'\gamma$, and the coefficient estimates suggest that this is linked to age and industries 2 (miscellaneous manufacturing) and 3 (retailing, etc). However, the R^2 is low, and heteroskedasticity is not a striking feature of the model.

6.3 Regression of y_1 on y_0 and z

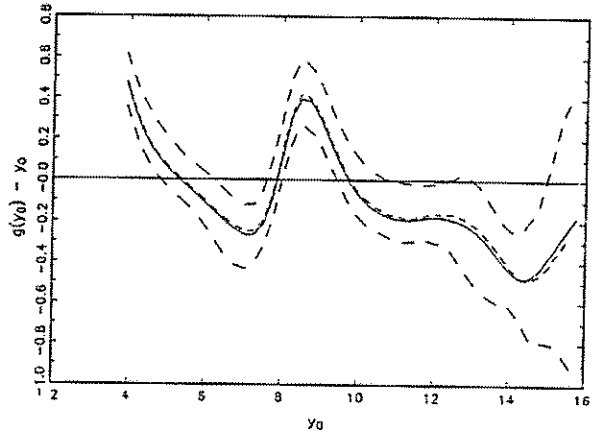
The final element required for construction of the joint distribution (5)-(6) is the density of y_1 conditional on y_0 and z. We estimate this first by using the additive semi-parametric regression estimator described in section 3.3 above. The results are reported in figure 6 and Table 9. They turn out to be very similar to the ordinary linear regression estimates of Table 5. The estimate of the function g(.) is shown in figure 6(a); it is close to linear, with a gradient slightly less than one. To illustrate the implications more clearly, we plot in figure 6(c) the function $g(y_0) - y_0$. Using the pointwise bootstrap confidence intervals as a guide to statistical significance, there is some evidence of a tendency for relatively rapid growth among small firms (with asset values of £0.05m - £0.15m) and also among medium-large firms (asset values in the range £3m - £13m). This is a form of nonlinearity that would be difficult to capture using conventional parametric specifications, although the departure from linearity is admittedly not large. This finding is reflected in the fact that R^2 is almost identical for the parametric and semi-parametric regressions.



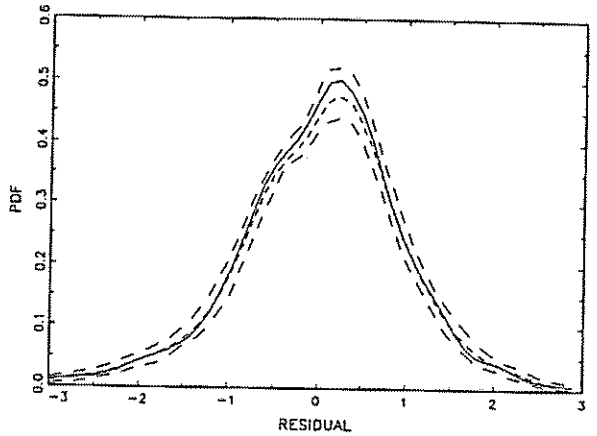
(a)



(b)



(c)



(d)

Figure 6 Surviving firms: additive semi-parametric growth regression
(____ = sample regression; _ _ _ = 90% confidence interval;
- - - = mean of bootstrap replications)

Table 9 Additive semi-parametric least squares estimation results (Standard errors computed from 40 bootstrap replications)

Covariate	Regression (dependent variable = log final size)		Semi-parametric skedasticity function (dependent variable = squared residual)	
	Coefficient	Standard error	Coefficient	Standard error
Log(age)	-0.047	0.021	1	-
Industry 1	0.046	0.052	0.265	0.237
Industry 2	0.118	0.053	0.285	0.457
Industry 3	0.056	0.056	0.270	0.444
Industry 4	0.188	0.070	0.380	0.326
Log(y0)	-	-	0.900	2.000
R^2	0.949		0.014	

Turning to the coefficient estimates in Table 9, we find very similar estimates to the regression coefficients in Table 5, although with slightly larger standard errors as one would expect. Growth is found to slow with the age of the firm and to be significantly faster than the norm in industries 1, 2 and 4 (engineering, etc., miscellaneous manufacturing and services).

We have chosen to investigate the role of heteroskedasticity through the following model:

$$var(y_1 \mid y_0, z) = \psi(z'\gamma + \gamma_0 y_0) \qquad (26)$$

with the first element of the vector γ normalised to unity. The function ψ and the parameters γ and γ_0 are then estimated by Ichimura's semi-parametric least squares regression technique. There is in fact no significant evidence of heteroskedasticity. The estimated function $\psi(.)$, plotted in figure 6(b) together with the pointwise bootstrap confidence interval, shows no significant departure from a constant value. The estimated coefficients presented in Table 9 have very large standard errors, and, indeed, are unidentifiable if the true function $\psi(.)$ is invariant. The distribution of the standardised residuals plotted in figure 6(d) is close to the normal.

8 The implied distribution of firm growth

The estimates presented above provide three alternative ways of constructing the joint distribution of firms size and death. This mixed discrete/continuous distribution is as follows:

$$f(\zeta = 1 \mid y_0, z) = \frac{f(y_0 \mid \zeta = 1, z)\, f(\zeta = 1 \mid z)}{f(y_0 \mid z)} \qquad (27)$$

$$f(y_1, \zeta = 0 \mid y_0, z) = f(y_1 \mid y_0, \zeta = 0, z)\, \left(1 - f(\zeta = 1 \mid y_0, z)\right) \qquad (28)$$

The three possibilities are:

(i) Estimate the three components of the right-hand side of (27) using the static normal-linear regressions and probit in columns 2-4 of Table 5 and the first component of the right-hand side of (28) using the regression in column 1 of Table 5.

(ii) As (i), except that the left-hand side of (27) is estimated directly, using the probit model in column 6 of Table 5.

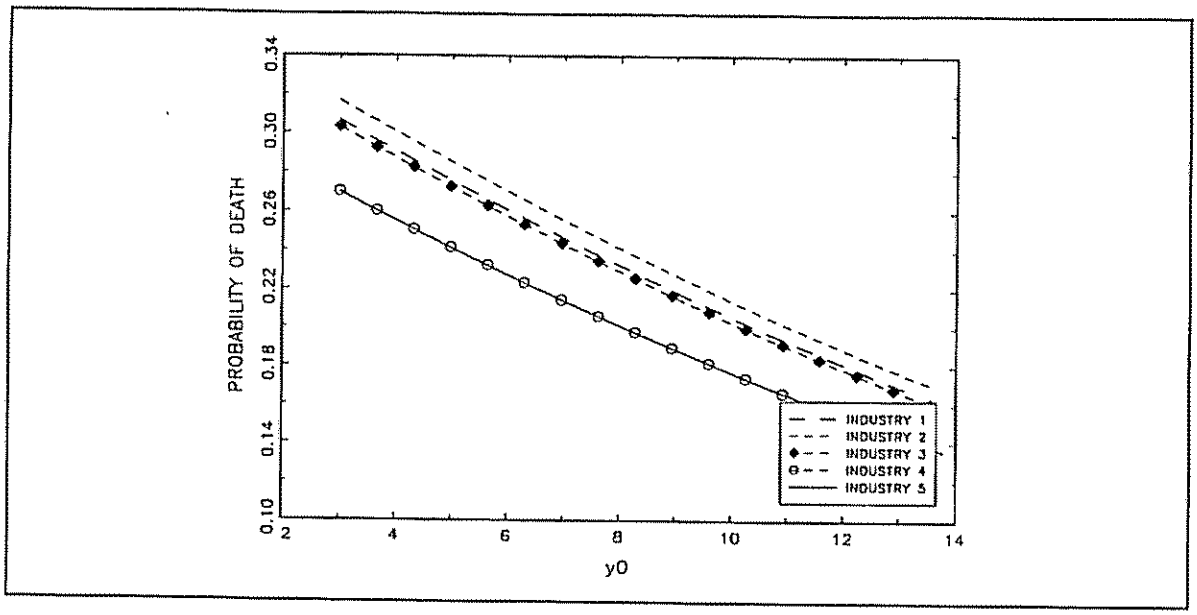
(iii) All components of the right-hand side of (27) and the first component of (28) are estimated semi-parametrically.

The three approaches entail an increasing degree of flexibility. Approach (i) constructs the distribution from a linear probit and a linear regression. Thus the estimated distribution depends on only two parameter vectors, estimated within the linear-homoskedastic-normal framework. Approach (ii) maintains the assumptions of linearity, homoskedasticity and normality, but introduces further flexibility through the decomposition of (27) into three components. Thus the estimated distribution involves four separately-estimated parameter vectors. Approach (iii) takes this a step further by relaxing the linearity, normality and homoskedasticity assumptions by means of semi-parametric

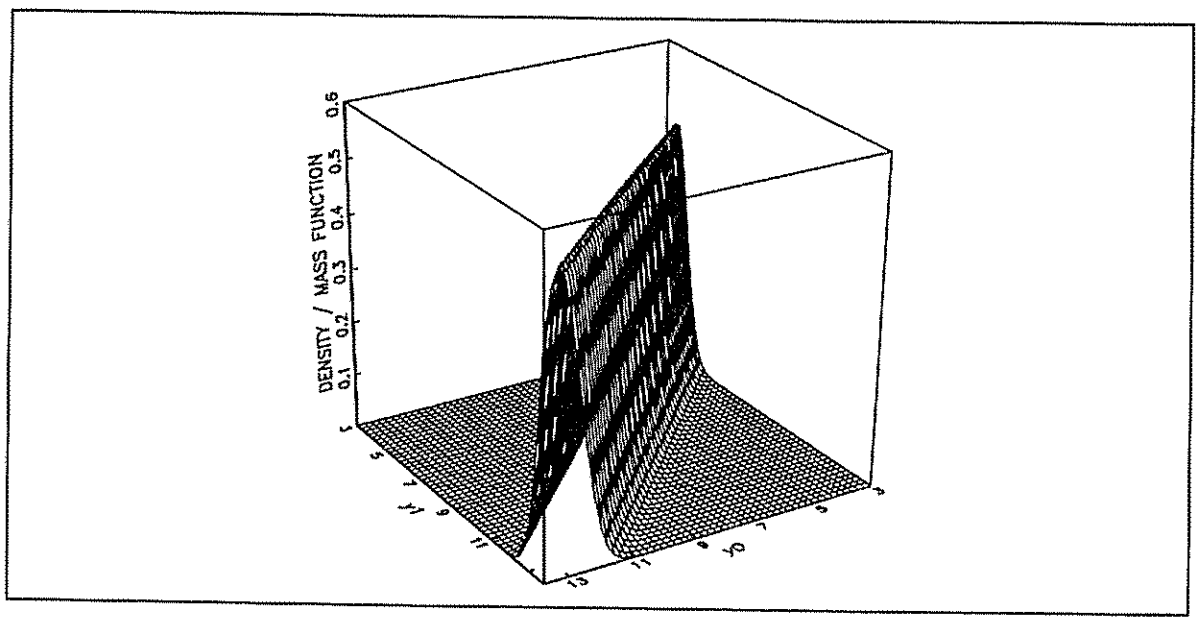
techniques. The three resulting estimates are plotted in Figures 7-9 below. The two parts of the distribution (27)-(28) are plotted separately. Expression (27) is plotted for each of the five industrial sectors, holding log age constant at the sample mean. Expression (28) is plotted as a surface in three-dimensions, with the four industry dummies set to zero and log age set to its sample mean.

The main drawback of the conventional linear regression-probit approach is clearly in the modelling of firm death. Probit estimation imposes a smooth monotonic relationship between the probability of death and initial size, with variations in the other explanatory variables producing moderate near-parallel shifts in the relationship. In Figure 8, some flexibility is introduced through the decomposition (27), and this allows non-parallel shifts in response to the explanatory variables. Thus, industry 4 is estimated to have a quite different schedule of size-specific mortality rates than industry 5. Taking this flexibility still further using the semi-parametric approach, Figure 9(a) suggests the existence of both heterogeneity across industries and non-monotonicity with respect to size. Indeed, using these more flexible methods, there appears to be little evidence of any simple relationship between firm mortality and size.

In contrast, conventional linear regression seems to describe the growth component of the distribution (expression (28)) pretty well. There is remarkably little difference between the plots in Figures 7(b) and 9(b), and the evidence of non-linearity in the conditional distribution of growth on size is confined to the level of fine detail apparent in the semi-parametric regression results presented in Figure 6(c) discussed above. Of course, this result may be partly an artefact of the remaining assumptions that have not been relaxed here. For example, the additivity assumption imposed on the semi-parametric growth-size regression (19) greatly reduces the formidable computing burden of these methods and improves statistical precision, but it also rules out possibly important interactions between size and age or industry.

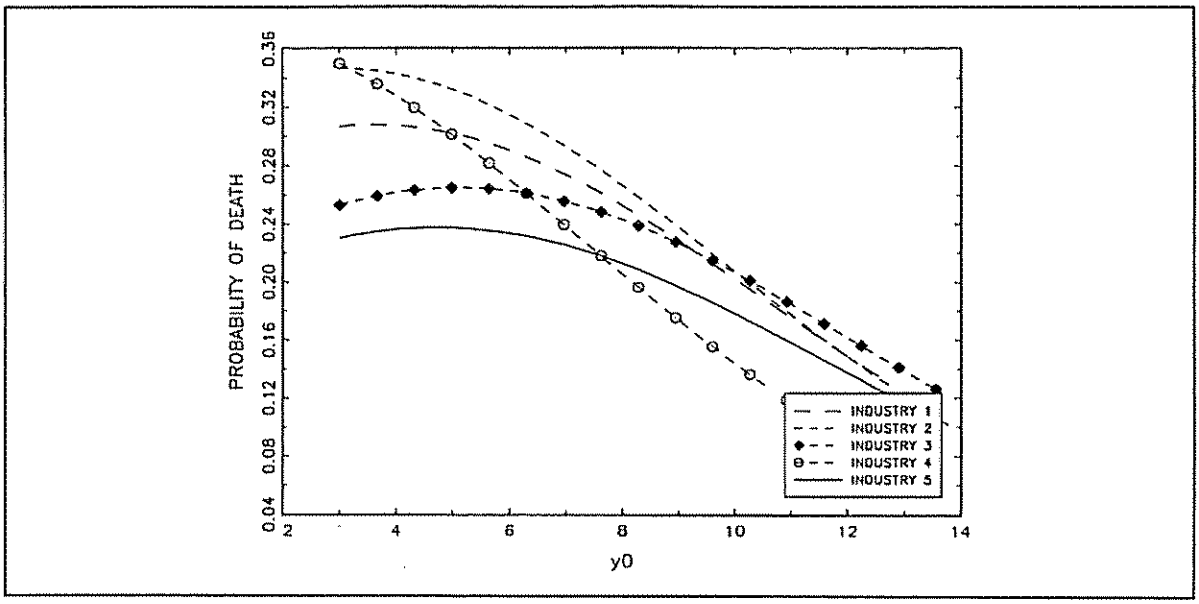


(A) $f(\zeta=1 \mid y_0, z)$

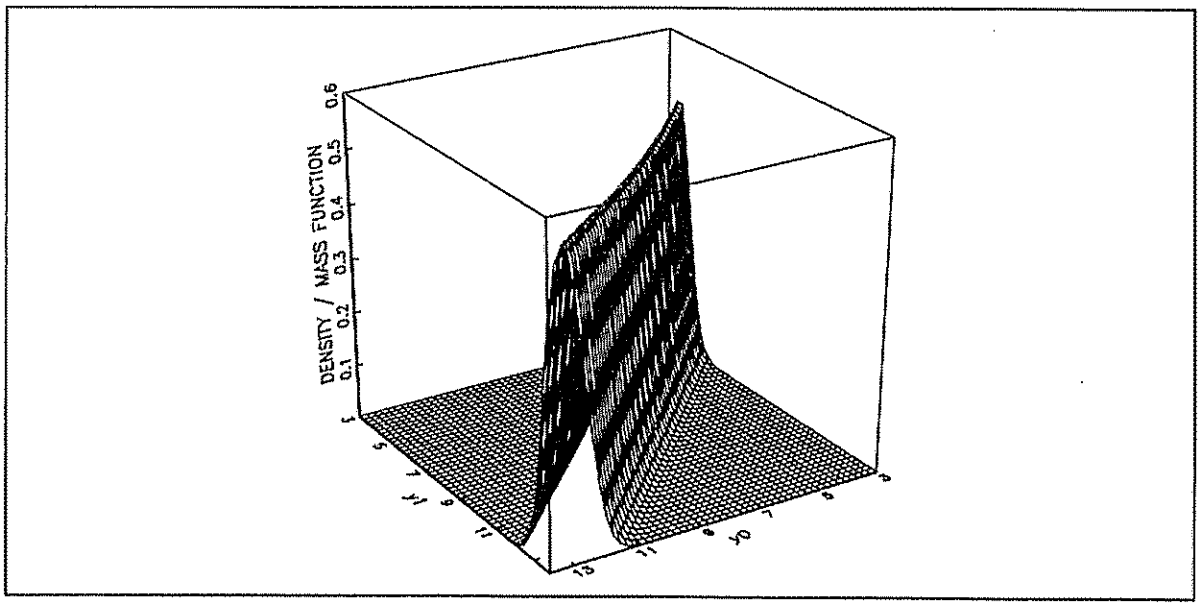


(B) $f(y_1, \zeta=0 \mid y_0, z)$

Figure 7 Linear-homoskedastic-normal model; $f(\zeta=1 \mid y_0, z)$ estimated directly

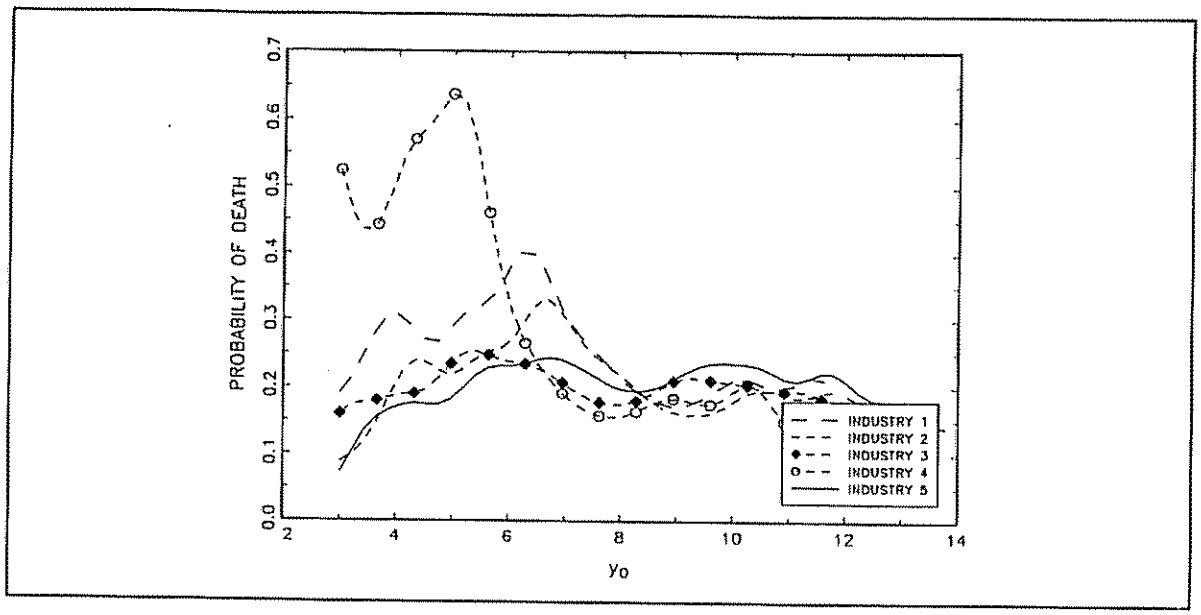


(A) $f(\zeta=1 \mid y_0, z)$

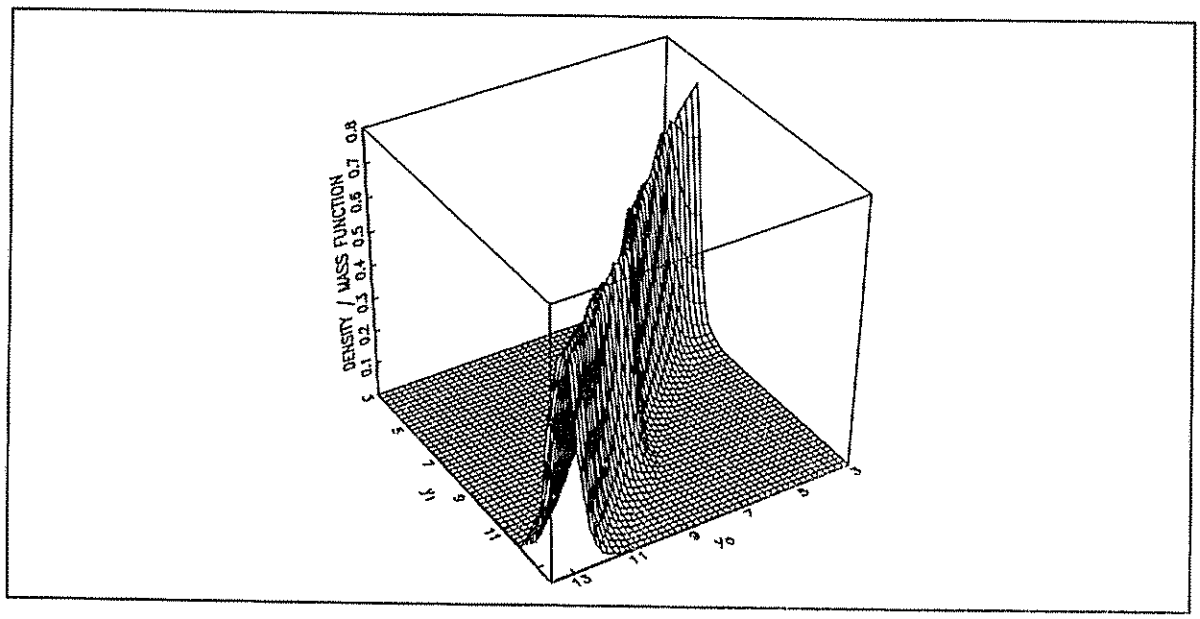


(B) $f(y_1, \zeta=0 \mid y_0, z)$

Figure 8 Linear-homoskedastic-normal model; $f(\zeta=1 \mid y_0, \mathbf{z})$ estimated indirectly



(A) $f(\zeta=1 \mid y_0, z)$



(B) $f(y_1, \zeta=0 \mid y_0, z)$

Figure 9 Semi-parametric model;
$f(\zeta=1 \mid y_0, \mathbf{z})$ estimated indirectly

9 Conclusions

Conventional assumptions of linearity, homoskedasticity and normality play important roles in applied econometric analyses of the process of firm evolution and mortality. In this study, we have attempted to investigate the validity of these assumptions in a simple exercise aimed at the estimation of the joint distribution of a discrete event (firm death within the period 1976-82) and a continuous outcome (1982 firm size) conditional on the 1976 size, age and sector of the firm. Our approach uses a decomposition of the conditional mortality rate into components which we estimate using kernel-based semi-parametric generalisations of linear regression and probit techniques, together with a partially-linear semi-parametric growth-size regression. Estimators of this type have been discussed by many econometric theorists and have been the subject of Monte Carlo simulations, but the number of applications to real data remains very small.

Applying this approach to a cross-section of companies covering a wide size range, we find that the the usual linear homoskedastic growth-size regression fitted to the subsample of survivors gives a reasonably good broad-brush description of the conditional distribution of growth rates, but that there is significant evidence of nonlinearity at the level of fine detail. Overall, we find little evidence of any relationship between growth and size, except for a tendency towards high growth among small firms (with asset values £0.05m-£0.15m) and medium-large firms (asset values (£3m-£13m). However, these nonlinearities, although apparently statistically significant, are far too fine to be detected by conventional specifications involving quadratic terms or size dummies.

In contrast, we find that the widely-used probit model of company mortality performs very poorly indeed. The use of a three-component decomposition of the conditional probability of firm death introduces additional separately-estimated parameter vectors, which greatly relax the restrictiveness of the linear probit model. More importantly, semi-parametric estimation relaxes the monotonicity property imposed by the probit model, and results in a quite different estimate of the mortality process. Whereas the probit model leads us to infer a strong negative relation between mortality and size (allowing for the effects of age and industry), the semi-parametric approach suggests a

non-monotonic relation rising to a peak for small-medium sized firms, then a flat or slightly declining schedule thereafter. The shape of this relation varies across industry, but the peak mortality rate is estimated to occur at a size corresponding to an asset value of around £0.5m in 1976.

Notes

1. The term probability density/mass function is henceforth abbreviated as pdf.

2. Alternatively, one could specify a function of this type for one or more of the quantiles of the conditional distribution. For example see Koenker and Bassett (1978) for a discussion of linear regression quantiles. Henceforth, we restrict attention to traditional regressions specified in terms of the expectation function.

3. This argument remains valid even if other causes are responsible for some deaths. Indeed, the existence of multiple possible causes of death makes the conventional selectivity model still less plausible, since the probit (24) is then attempting to approximate a probability that would involve the parameters of all the separate cause-of-death processes.

References

Arellano, M. and S.Bond (1991). Some tests of specification for panel data:Monte-Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277-297.

Breiman, L., W. Meisel and E. Purcell (1977). Variable kernel estimates of multivariate densities. *Technometrics* 19, 135-144.

Cosh, A.D., Hughes, A. and Kambhampati, U. (1993). Size, Growth and Failure: An Analysis of the UK Quoted and Unquoted Company Sectors. Small Business Research Centre *Working Paper* 32, University of Cambridge.

Cosh, A.D. and Hughes, A. (1994). Acquisition Activity in A. Hughes and D.J. Storey (eds) *Finance and the Small Firm*, Routledge, London.

Cosh, A.D. and Hughes, A. (1995). Failures, Acquisitions and Post Merger Success. The Comparative Financial Characteristics of Large and Small Companies. ESRC Centre for Business Research *Working Paper* 18, University of Cambridge.

Cosslett, S.R. (1983). A distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51, 765-782.

Davies, S.W. and B.R. Lyons (1982). Seller concentration: the technological explanation and demand uncertainty. *Economic Journal* 92, 903-919.

Deaton, A.S. (1989). Rice prices and income distribution in Thailand: a non-parametric analysis. *Economic Journal* 99, 1-37.

Dunne, J.P. and A. Hughes (1994). Age, size, growth and survival: UK companies in the 1980s. *Journal of Industrial Economics* 42, 115-140.

Evans, D.S. (1987a). Tests of Alternative Theories of Firm Growth. *Journal of Political Economy* 95(4), 657-674.

Evans, D.S. (1987b). The Relationship between Firm Growth, Size and Age: Estimates for 100 Manufacturing Industries. *Journal of Industrial Economics* (35), 567-581.

Gabler, S., F. Laisney, and M. Lechner (1993). Seminonparametric estimation of binary-choice models with an application to labor-force participation, *Journal of Business and Economic Statistics* 11, 61-80.

Gallant, A.R. and D.W. Nychka (1987). Semi-nonparametric Maximum-likelihood Estimation. *Econometrica* 55, 363-390.

Hall, B.H. (1987). The Relationship between Firm Size and Firm Growth in the Manufacturing Sector. *Journal of Industrial Economics* 35(4), 583-606.

Han, A.K. (1987). Non-parametric analysis of a generalised regression model. The maximum rank correlation estimator. *Journal of Econometrics* 35, 303-316.

Hannah, L. and J.A. Kay (1977). *Concentration in Modern Industry*. London: Macmillan.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hart, P.E. (1965). *Studies in Profit, Business Saving, and Investment in the United Kindom 1920-62: Vol. 1*. London:George Allen and Unwin.

Hausman, J.A. and D.A. Wise (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment, *Econometrica* 47, 455-473.

Heckman, J.J. (1974). Shadow prices, market wages and labour supply, *Econometrica* 42, 679-694.

Horowitz, J.L. (1992). A smoothed maximum score estimator for the binary

response model. *Econometrica* 60, 505-531.

Horowitz, J.L. (1993). Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics* 58, 49-70.

Hughes, A. (1993). Mergers and Economic Performance in the UK: A Survey of the Empirical Evidence 1950-1990 in M. Bishop and J.A. Kay (eds) *European Mergers and Merger Policy*, Oxford University Press, Oxford.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71-120.

Klein, R.W. and R.H. Spady (1993). An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica*. 61, 387-421.

Knight, J. (1987). Top Companies in the DTI Company Accounts Analysis, *Statistical News* 76, February, 16-19

Koenker, R. and G. Bassett (1978). Regression quantiles, *Econometrica* 46, 33-50.

Kumar, M.S. (1984). *Growth, Acquisition, and Investment.* Cambridge:Cambridge University Press.

Lewis, C. (1979). Constructing a Sampling Frame of Industrial and Commercial Companies *Statistical News* 44, February, 6-11.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 27, 313-334.

Nickell, S.J. (1981). Biases in dynamic models with fixed effects. *Econometrica* 49, 1417-1426.

Prais, S.J. (1976). *The Evolution of Giant Firms in the UK.* Cambridge: Cambridge University Press.

Pudney, S.E. (1993). Income and wealth inequality and the life cycle. A non-parametric analysis for China. *Journal of Applied Econometrics* 8, 249-276.

Quah, D. (1993). Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics* 95, 427-443.

Ramsey, J.B. (1969). Tests for Specification Errors in Classical Linear Squares Regression Analysis. *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica* 56, 931-954.

Samuels, J.M. (1965). Size and the Growth of Firms. *Review of Economic Studies* (32), 105-112.

Samuels, J.M., and A.D. Chesher (1972). Growth, Survival, and the Size of Companies 1960-69. in Cowling, K.(ed.) *Market Structure and Corporate Behaviour*. London: Grey-Mills.

Schmalensee, R. (1989). Inter-Industy Studies of Structure and Performance. in Schmalensee, R. and R. Willig (ed.). *Handbook of Industrial Organisation, Vol.2*. North-Holland:Amsterdam.

Singh, A. and G. Whittington (1968). *Growth, Profitability, and Valuation*. Cambridge: Cambridge University Press.

Singh, A. and G. Whittington (1975). The Size and Growth of Firms. *Review of Economic Studies*, 15-26.

Storey, D., K. Keasey, R. Watson, and P. Wynarczyk (1987). *The Performance of Small Firms*. Routledge: London.

Appendix

Definition of Industries

Industry 1 Metal, Mechanical & Instrument Engineering, Electrical & Electronic Engineering; Office Machinery etc., Shipbuilding, Vehicles, Metal goods.

Industry 2 Food, Drink, Tobacco, Chemicals and Man-Made Fibres, Textiles, Leather and Leather Goods, Footwear; Clothing, Non Metallic Mineral Products, Timber; Furniture, Paper; Printing; Publishing, Other Manufacturing, Mixed Activities in Manufacturing: companies with over 50% of activity in manufacturing and which are engaged in 3 or more activities none of which account for 40% or more of activity.

Industry 3 Wholesaling (other than petroleum products), Retailing; Hotels and Catering; Repair of consumer goods and vehicles.

Industry 4 Business services; leasing, other services, mixed activities in non manufacturing: companies with over 50% activity in non manufacturing and which are engaged in 3 or more activities non of which account for 40% or more of activity.

Industry 5 Agriculture; Fishing and Forestry, Mineral and Ore Extraction, Oil, Construction, Transport and Communication.