Small Bus Econ (2011) 37:155–165
DOI 10.1007/s11187-009-9240-4

Small business survival and sample selection bias

Hanas A. Cader · John C. Leatherman

Accepted: 12 October 2009 / Published online: 7 November 2009

Abstract Analyses of small business and the factors affecting their survival are fairly common in the research literature. The level of research interest may stem from the fact that in the US, only about half of all new small businesses survive after 4 years (Headd 2003). However, research attempting to understand the phenomenon that employs data using only information from and about surviving firms may lead to erroneous conclusions regarding the factors that influence firm survival and failure. In this paper, we provide evidence that omitted information about the firms that disappear from the research data over time leads to biased coefficient estimates. Comparing the Heckman two-step estimation approach of switching regression models to a semi-parametric Cox hazard model, the Accelerated Failure Time (AFT) model, we conclude that the Cox ATF approach is the most appropriate model for firm survival analysis.

Keywords Firm survival · Omitted observation · Selection bias · Two-stage estimation

JEL Classifications C01 · C25 · C52 · L26

H. A. Cader (✉)
American University of Kuwait, Safat, Kuwait
e-mail: mcader@auk.edu.kw

J. C. Leatherman
Kansas State University, Manhattan, KS, USA

1 Introduction

Business survival depends on a multiplicity of factors, including firm, industry, regional, and national economic conditions. Earlier research suggested that firm entry and exit rates were very much correlated to overall industry conditions (Dunne and Roberts 1991), where high exit rates were associated high entry rates. Others suggested that firm entry and exit were more closely associated with regional economic conditions (Keeble and Walker 1994; Reynolds et al. 1994). While a considerable body of research investigated the factors that influence firm entry, exit, and survival, Audretsch et al. (2001, p. 10) suggested that, "…there is a lack in knowledge concerning questions like who enters and exits, what determines this mobility and what are its effects, in particular on economic performance." While some focused exclusively on the factors that determine entry (Orr 1974; Geroski 1991; Baldwin 1995; Anagnostaki and Louri 1995), others have focused on exit (Marcus 1967; Audretsch and Mata 1995; Doi 1999). Recently, a few researchers considered entry and exit simultaneously (Arauzo et al. 2007; Kleijweg and Lever 1996; Lay 2003), building on the hypothesis proposed by Shapiro and Khemani (1987). The common proposition of these studies is that firm entry and exit are related (symmetry), and the factors that affect entry also are likely to affect the exit or vice versa (simultaneity).

The response variables employed in firm survival analysis are surviving firms and the firms that exited.

Inferences based on models of firm entry, exit, survival, or entry and exit together can be misleading due to self-selection-bias in much of the extant research. Three types of self-selection-bias can be identified in model specification. They include: (1) firm entry analysis that failed to capture the firms that did not enter; (2) in survival analysis, the exclusion of exited firms; and (3) as firms exit, the analysis failed to incorporate surviving firms. The relative degree of the selection-bias problem is higher for small business analyses than for the study of larger businesses. In an economy such as the US, not only are large numbers of small firms created every year, a considerable number of small firms exit the market as well. Previous research has found that small firms have high probability of failure (Dobrev 2001; Mitchell 1994; Reid 1995). Methodological issues related to self-selection bias in small business research are not trivial insofar as policy makers view the creation of new small businesses as key economic development strategy in urban and rural places alike (Shaffer et al. 2005).

The magnitude of the selection bias depends on the influence of a given explanatory variable on the response variable (survival or exit) (Heckman 1979). For example, if the explanatory variable has a strong impact on survival, the estimated coefficient is subject to a downward bias as the inclusion of exiting firms would have lowered the value of the coefficient estimate (Storey 2000). Another problem associated with selection bias is that it influences the level of significance of the variables, where an insignificant variable could become significant or vise-versa in the correct model specification. Policy recommendations often are based on the direction and the magnitude of the estimated coefficients, which may result in the adoption and implementation of erroneous economic development policies. In such cases, the selection bias problem is likely to have far-reaching implications for firm creation, survival, and death.

Heckman (1990) suggested that switching regression models can solve the sample selection-bias problem. The application of either Heckman's two-step method or the Expectation Maximization (EM) algorithm appears to have been rarely used in firm-survival modeling. However, the switching regression model approach only examines the sample-bias problem and fails to capture the explanatory factors influencing firm entry and exit. The Cox proportional hazard regression model (Cox 1972) is frequently used in survival and death analysis. This is one of a variety of proportional hazard models that have been developed. These models typically depend on the distribution of the baseline hazard (hazard probability) and its association with a series of explanatory variables (Jain and Kini 2000). If the baseline hazard function is not known, then semi-parametric Cox hazard models are employed in the analysis (Cleves et al. 2004). The Accelerated Failure Time (AFT) model is one form of the semi-parametric Cox hazard model and has been used recently to examine the impact of various explanatory variables in firm death or hazard probability.

The objectives of this study are to determine the presence of sample selection bias in firm-survival analysis resulting from the exclusion of information about firms that cease to operate and to examine the impact of a variety of explanatory variables on firm death (hazard) and survival. We apply the Heckman two-step estimation method to small-firm entry in the State of Kansas, USA. This is followed by the application of the Cox AFT model to examine the factors that influence small-firm survival. The contribution of this study is empirical in that it sheds light on the biased-coefficient estimate problem in firm survival analysis. The remaining sections of this paper include five sections. The next section presents a review of the literature, followed by the econometric methodology used to overcome the problem of missing information. We then discuss the data. The final sections present the results and conclusions regarding the preferred analytical approach for firm survival analysis.

2 Literature review

Organizational life-cycle theory suggests that any firm coming into existence proceeds through multiple stages. These stages are similar to the growth of biological organisms. The stages include birth, survival or expansion, consolidation, diversification, and finally decline and death (Hanks et al. 1993). Any firm has two definite stages: start-up or birth (Tichy 1980) and the eventual exit or death (Kimberly and Miles 1980). In between, there are stages in which the firm tends to change its size (Mintzberg 1979, 1989).

The first stage is commonly known as entrepreneurial stage (Quinn and Cameron 1983), extensively studied in the entrepreneurship literature. The entrepreneurial assessment of new firm creation stems from the argument that firm start-up is largely attributable to the personal characteristics of the individual proprietor (Peteraf and Shanley 1997; Reuber and Fischer 1999; Praag 2003). Among all stages, the survival stage is most crucial. Churchill and Lewis (1983) found that firms tend to stagnate in this stage. As a result, many firms eventually cease to operate.

Many studies have focused on the survival of firms. The factors that affect survival can be broadly classified into those that are specific to the firm (e.g. size, type), entrepreneur (age, education, etc.), industry (manufacturing, technology-based), region (metro, non-metro), or the combination of these factors. In a recent study, Strotmann (2007) found that sector-specific conditions are conducive to firm survival, while agglomeration economies are not. Using Swedish firm-level data, Persson (2004) showed that firm survival increases with age and the size of the firm, as well as the educational attainment of the employer. The size of the firm has been shown to be an important determinant of firm survival, as the ability to attract financial capital increases with firm size (Audretsch 1991). Investigating the influence of a region's human capital stock on firm survival, Acs et al. (2007) found a negative relationship between the high school dropout rate and new firm survival in the service sector.

In survival analysis model specification, the response variable is the surviving firm (number of firms or the odds ratio of surviving firms), excluding firms that ceased to operate. However, the conditions that influenced some firms' survival were the very ones caused the death of others. The parameter estimates from models that excluded exited firms tend to be biased as a result of the omitted information. Audretsch and Thurik (1999) identified two types of selection-bias problems in manufacturing firm survival analysis. The first is the over-representation of larger firms. The other is growth that is observed in surviving firms. Econometrically, the first source of bias is considered to be less problematic relative to the second. In the former case, no information is missing. There is merely an under-representation of smaller firms. In the latter case, the information loss is apparent and purposeful. Simple establishment counts are used without regard to changing employment levels. Missing information has serious implications for the resulting coefficient estimates because it omits an entire class of firms in the analysis. For example, Thompson (2005) found that missing information related to firm-owner's experience resulted in an upward bias of coefficient estimates.

The sample-selection problem is likely to be more severe in small-firm survival analysis as a larger percentage of the smaller firms exit within the first year of operation. Econometrically, missing data are handled using mechanisms to estimate the missing values. Little and Rubin (2002, p. 12) discuss three kinds of missing-value mechanisms. A complete dataset is defined as $Y = (y_{ij})$ and a missing-data indicator matrix $M = (\mathrm{m}_{ij})$. Complete data Y consists of observed data (Y_{obs}) and missing data (Y_{mis}).

1. Missing Completely at Random (MCAR): the probability of missing data does not depend on the values of the data Y.
2. Missing at Random (MAR): the probability of missing data depends only on the component of Y_{obs} of Y.
3. Not Missing at Random (NMAR): the probability of missing data depends only on the component of Y_{mis} of Y.

A full discussion of the estimation techniques for these three missing-data mechanisms can be found in Little and Rubin (2002). Since the NMAR missing data mechanism is intuitively appealing, there are many approaches available to solve this category of missing values. The most common solutions are Heckman's (1976) two-step method and the Expectation Maximization (EM) algorithm (Dempster et al. 1977). It appears that most of the empirical work on Heckman's two-step application, however, focused on wages and labor force participation, following the original work by Heckman (1976). Chakravorty (2003) noted that a good survey of Heckman's two-step application is found in Vella (1998) and Winship and Mare (1992).

Examining the pre-post world reforms period industrial location choice decisions in India, Chakravorty (2003) found that OLS/Heckman selection models were robust, but less successful in explaining the variation in the distribution of new investments. Acs (2002) used the Heckman two-step estimation procedure to examine the impacts of research and development (R&D), innovation, and wages on high-technology employment in 36 Metropolitan Statistical Areas (MSAs). Low employment in high-tech industries contributed to the selection bias problem. While the ordinary least square (OLS) estimates were upward biased, the coefficient for selection bias (Mills ratio) was insignificant. Costa et al. (2004) studied the location choice of new firms in five sectors using Spanish establishment data. New firm creation in an industry was explained by externalities (diversity and specialization), population density, and population size (more than 500,000, 100,000, etc.). The Heckman two-step method was used to estimate the unbiased parameters, and the Mills ratio (to test for selection bias) was significant in six models and insignificant in four others (Costa et al. 2004). In this research, the Heckman two-step method was applied to small firms in three industries. The industries include the goods-producing industry, the service-producing industry, and the information technology (IT)-producing industry. We hypothesize that relatively lower levels of employment in the IT-producing industries will lead to a more severe selection-bias problem for this sector compared to goods- and service-producing industries.

3 Methodology

A general switching model is presented here using Heckman (1990, p. 313) and Vella (1998, p. 129). X, Y and Z are the variables of interest and lower case (x_i, y_i, and z_i) refer to observations of that variable.

$$y_i^* = x_i^{'}\beta + \varepsilon_i; \quad i = 1, 2, \ldots, N, \tag{1}$$

$$d_i^* = z_i^{'}\gamma + v_i; \quad i = 1, 2, \ldots, N, \tag{2}$$

$$d_i = 1 \quad \text{if } d_i^* > 0\,;\ d_i = 0, \text{ otherwise} \tag{3}$$

$$y_i = y_i^* * d_i \tag{4}$$

where y_i^* is the latent endogenous variable with observed counterpart y_i, β and γ are coefficients to be estimated, and ε_i and v_i are zero mean error terms uncorrelated with regressors, but correlated with each other ($E[\varepsilon_i|v_i] \neq 0$).

The variable Y is industry employment. Current year non-zero industry employment is observed in a particular region. However, the employment of establishments that exited (death) or migrated in the previous year is not observed in the current year. Since it is possible to track establishment data in previous years, establishment employment for those firms are recorded as zero in the current year. Variable Y is partitioned into observations (Y_1) that are greater than zero and equal to zero (Y_2). The observations are defined as y_{1i} and y_{2i}. If a firm did not survive in the subsequent year, the firm is considered to be dead during the current year ($y_{2i} = 0$) and otherwise ($y_{1i} = 1$). A restrictive form of the general model is:

$$y_{2i} = \beta_2^{'} x_{2i} + \varepsilon_{2i}, \tag{5}$$

$$y_{1i} = \beta_1^{'} x_{1i} + \varepsilon_{1i} \quad \text{if } y_{2i} > 0 \tag{6}$$

where the error terms ε_{1i} and ε_{2i} have a zero mean with constant variance. While the parameter of interest (β_1) can be estimated using (6), the estimates are potentially biased because of the omitted-variable problem. Although all the X_1 variables are found in X_2, there needs to be at least one variable that is different from X_1 in X_2. Empirically, this assumption has often been ignored. After specifying conditional density $f(Y|X,\beta)$, the following equation can be derived using Heckman (1979, p. 156):

$$Y_i = \beta^{'} X_i + \mu \hat{\lambda}_i + \eta_i \tag{7}$$

where λ_i is referred as the Mill's ratio and a monotone decreasing function of the probability that an observation is selected into the sample. The β is a consistent parameter estimate using OLS. Sample-selectivity bias can be proven by testing $\mu = 0$ and $E(\eta) = 0$. If these conditions hold, then the null hypothesis (no selection bias) is not rejected (Vella 1998, p. 134).

In the AFT model, the hazard function ($h(t|X_i)$ is defined as the product of vector covariates and the baseline hazard function. Cox (1972) assumed the vector covariates function as an exponential function $\left(\exp(\delta^{'} X_i)\right)$, implying that $(h(t|X_i) \geq 0$. In the AFT model, the baseline hazard ($h_0(t)$) is unspecified. The advantage of un-specifying the hazard function is the reduction of potential errors of incorrect specifications (Cleves et al. 2004). The logarithmic transformation of left and right hand variables results the following hazard regression model:

$$\ln(h(t)) = \delta^{'} X_i + \varpi$$

where ϖ is the error term with i.i.d. The ES hazard time is measured as the time difference between

December 31 of the exit year and the reported initial liability (start) date found in the Kansas Quarterly Census of Employment and Wages data used in this analysis.

4 Data

The primary data used in this analysis were obtained from Kansas Quarterly Census of Employment and Wages (QCEW) data files from 1990 to 2003 for all 105 Kansas counties. Nationally, QCEW data consist of 98% of non-farm payroll employment (Knaup and Piazza 2007). QCEW data are used to track employers whose employees are subject to unemployment compensation insurance taxes. All employers with employees are subject to reporting requirements. These were fully disclosed firm-level annual employment/establishment data for the 13 years included in the analysis. While the US Small Business Administration typically defines a small business as one employing 500 or fewer workers (SBA 2008), for purposes of this research small businesses were defined as all establishments with 100 or fewer employees. While this definition is debatable, previous studies have shown that firms with less than 100 employees constitute about 90% of the small and medium-sized enterprises compared to the SBA definition (Santarelli and D'Altri 2003).

Apart from firm deaths, firms also are likely to grow beyond the 100-employee threshold. A firm in 1 year may have fewer than 100 employees and more than 100 the next. As the total number of employees exceeds 100, the firm was dropped from the analysis. However, if a firm goes out of business (firm death), its data were retained for purposes of the analysis.

The data used to explain small firm survival were obtained from a variety of sources. For example, industry classification data were obtained from US Department of Commerce (DOC) publications, while other data came from Woods and Poole Complete Economic and Demographic Data Source (CEDDS, 2006), US Geological Survey, National Science Foundation, IMPLAN (Impact Analysis for Planning 1999) software, the Office of Management and Budget (OMB 2000), USDA Economic Research Service, and County Business Patterns. All data items, definitions, and sources are shown in Table 1.

Goods-producing industries include manufacturing, construction, and agricultural services; service-producing industries include services, retail trade, wholesale trade, finance, insurance, and real estate, and transportation, communication and public utilities; the IT-producing industry is defined based on DOC criteria. The DOC identified intensive producers of IT as industries that (1) produce, process, or transmit information goods or services as either intermediate or final products, or (2) provide the necessary infrastructure for the Internet (Henry et al. 1999). Thirty (30) four-digit SIC industries were selected as information-technology producers based on the DOC criteria.

5 Results

Before estimating the selection bias, we present the magnitude of the firm deaths for the first 5 years following the firm's entry. The QCEW data have a unique employer identification number, permitting the tracking of individual firms. Thus, our analysis is able to determine whether a unique firm did, in fact, disappear from the data. While the data from 1990 to 2003 were available, a firm's survival up to 5 years following entry limits the use of data for firms created after 1998. Firms created after 1998 were unable to fulfill the 5-year survival criteria. Table 2 presents the relative frequency of all firm deaths during the first 5 years following entry.[1] Of the 90,134 observations, 37,937 survived after 5 years. About 15% of the firms ceased operation before the end of the first year, while another 13% exited the market before the end of the second year. More than 40% of the firms did not survive after 3 years, and 42% of the firms survived after 5 years. Our findings are consisten with Knaup and Piazza (2007) who studied new establishment entry for a 7-year period using the US national QCEW data and found that about 40% of the firms survived after 5 years.

The comparative death analysis among the three industry sectors (Tables 3, 4, 5) indicates that there were more deaths observed in the IT-producing

[1] For context, recall that in 1991, the US was climbing out of recession. The period through the mid- to late-1990s was one of strong national economic growth. In the early 2000s, the national economy began to slump once again.

Table 1 Definition of variables and data sources incorporated in the research analysis

Descriptor	Dependent variables	Source
Employment	Annual average counts	QCEW data
Establishments	Annual establishment counts	QCEW data
	Independent variables	
Density	Population density: population per square mile	Woods and Poole Economics, Inc.
Labor quality	Quality of the labor force: percent of employees in knowledge industries	QCEW data and Beck (1992)
County employment growth rate	County employment growth in percentage	QCEW data
Highway	Presence of interstate highway in the county	Kansas State Highway Map
Clustering	Industry clustering: location quotient (LQ) $\mathrm{LQ} = \frac{\left(\frac{\text{County industry employment}}{\text{Total county employment}}\right)}{\left(\frac{\text{National industry employment}}{\text{Total national employment}}\right)}$	QCEW data and County Business Patterns
Integration	Vertical integration: indirect output multiplier in million dollars	IMPLAN software (MIG Inc. 1999)
Intensity	Labor intensity: direct employment multipliers in million dollars per employment	IMPLAN software

industry within the first year and within the first 5 years. About 39.7% of the IT-producing firms survived after 5 years. Service-producing industry firms were most likely to survive after 1 year and 5 years. These results suggest that the environment within which IT-producing firms function is more volatile than for goods-producing firms and least volatile for the service-producing firms. These results also are consistent with Knaup and Piazza (2007) who found that the information sector had the lowest survival rate, while the service-producing firms had the highest survival rate.

If an analysis focused exclusively on surviving firms, a substantial number of firm deaths would be ignored. Focusing solely on surviving firms allows insight into the influence that certain factors or conditions have on survival, but offers nothing relative to how these same factors also foster corollary deaths.

Table 2 Relative frequency of all firm deaths

Survival in years	Count	Relative frequency (%)	Cumulative relative frequency (%)
<1	13,387	14.9	14.9
<2	11,474	12.7	27.6
<3	11,635	12.9	40.5
<4	8,743	9.7	50.2
<5	6,958	7.7	57.9

Source: Authors' estimates using QCEW data

5.1 Results for selection bias correction

The selection bias estimation and correction involve three steps. The first step is the estimation of an Ordinary Least Square (OLS) model including latent variables. A probit model is then estimated in the second stage to calculate the inverse Mill's ratio. The third stage involves re-estimating the OLS model and including the inverse Mill's ratio as an explanatory variable. The level of significance of the inverse Mill's ratio coefficient was examined for the presence of the selection bias problem. The size and level of significance of the Mills ratio reveal the relative severity of the selection bias problem. The results from the three estimates are presented in Table 6. While the coefficients for the first OLS model and probit model are not important in determining the selection bias problem, the results are presented for comparison.

The OLS and Heckman two-step models were used to determine the factors that were important for small firm survival and growth. Compared to the OLS model, the direction and magnitude of the coefficients

Table 3 Relative frequency of IT-producing industry firm deaths

Survival in years	Count	Relative frequency (%)	Cumulative relative frequency (%)
<1	426	17.0	17.0
<2	219	8.7	25.7
<3	334	13.3	39.0
<4	302	12.0	51.0
<5	232	9.2	60.3

Source: Authors' estimates using QCEW data

Table 4 Relative frequency of goods-producing firm deaths

Survival in years	Count	Relative frequency (%)	Cumulative relative frequency (%)
<1	2,462	16.5	16.5
<2	1,659	11.1	27.7
<3	1,986	13.3	41.0
<4	1,519	10.2	51.2
<5	1,105	7.4	58.7

Source: Authors' estimates using QCEW data

Table 5 Relative frequency of service-producing firm deaths

Survival in years	Count	Relative frequency (%)	Cumulative relative frequency (%)
<1	10,471	14.6	14.6
<2	9,108	12.7	27.2
<3	9,398	13.1	40.3
<4	7,002	9.7	50.1
<5	5,616	7.8	57.9

Source: Authors' estimates using QCEW data

and the *r*-square values in the two-step OLS model differ considerably. For example, in the IT-producing industries, lagged employment had a significant positive impact on firm survival in both the OLS and two-step OLS models. However, the value of the coefficient in the two-step OLS model was one-third of the value in the OLS model. The population density had a significant impact in both the OLS and two-step OLS models, but the magnitude of its impact was larger in the two-step model. Vertical integration had a positive and significant impact in the OLS model, but was negative and insignificant in the two-stage model. Presence of an interstate highway had a positive but insignificant impact on the IT-producing industry in the OLS model, but a significant negative impact in the two-step model.

The OLS and two-step models show a similar trend in the goods-producing industries, with the exception that county employment growth was positive and significant in OLS model while showing no effect in the two-step model. Further, labor intensity was negative and significant in the OLS model, but had almost no impact in the two-step model. The service industry data analysis showed a pattern similar to the results for the goods-producing industry, except that vertical integration had a significant negative impact in the OLS model, while being positive and significant in the two-step model.

The *r*-square value was consistently higher in original OLS models for all three industries. Population density had similar impacts in both the OLS and two-step models for all three industries. Most importantly, the coefficient for the inverse Mill's ratio (lamda) is negative and significant in all three models, implying the presence of a selection bias problem in firm survival analysis in all three industries. These results suggest that the two-step OLS estimates yield unbiased estimators for firm survival.

Considering the two-step OLS models across all industries, lagged employment, population density, and county employment growth had a significant positive impact on firm survival in the IT-producing industry. The lagged employment coefficient significance indicates that larger firms in the IT-producing industry were more likely to survive. Population density and the lagged employment variables also had significant positive impacts on the survival of goods-producing industry firms, but the magnitude of the population density coefficient was close to zero. A similar result was observed in the service-producing industry firm analysis, except that vertical integration also had a significant positive impact on survival. By far, vertical integration in the service-producing industry was the largest significant coefficient estimated using the two-step OLS method in all three industries. The level of significance of the vertical integration variable makes intuitive sense in that input supply from other industries is important to firm survival. Similarly, backward integration likely reduces transaction costs (Grossman and Hart 1986), uncertainty in production volume (Walker and Weber 1984), and the presence of beneficial market competition (Walker and Weber 1987). A more populous

Table 6 Estimated coefficients for IT-producing, goods-producing, and service-producing industries

Variables	IT-producing industry			Goods-producing industry			Service-producing industry		
	OLS	Probit	2 stage OLS	OLS	Probit	2 stage OLS	OLS	Probit	2 stage OLS
Intercept	0.79*	1.78*	28.09*	0.54*	1.62*	13.65*	0.46	1.61*	15.23*
	(0.35)	(0.04)	(1.48)	(0.12)	(0.01)	(0.27)	(0.06)	(0.01)	(0.15)
Lag employment	0.93*	0.01*	0.33*	0.94*	0.02*	0.53*	0.95	0.02*	0.53*
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Population density	0.00*	0.00*	0.02*	0.00*	0.00*	0.00*	0.00	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
County employment growth	0.06*		0.06*	0.01*		0.00	0.02		0.01*
	(0.01)		(0.01)	(0.00)		(0.00)	(0.00)		(0.00)
Highway	0.21		−0.49*	0.04		0.03	0.09		0.01
	(0.15)		(0.19)	(0.05)		(0.06)	(0.02)		(0.03)
Quality of labor force	-0.02		−0.01	0		-0.01	0.00*		0.00
	(0.01)		(0.01)	(0.00)		(0.00)	(0.00)		(0.00)
Labor intensity	0.13		−0.02	-0.01*		0.00	-0.48*		−0.34
	(0.08)		(0.09)	(0.00)		(0.00)	(0.23)		(0.3)
Industry cluster	0.07		−0.21	0.05		0.02	0.08		0.08
	(0.18)		(0.22)	(0.04)		(0.05)	(0.04)		(0.06)
Vertical integration	2.76*		−1.24	0.29		1.46	-0.87*		1.4*
	(1.25)		(1.51)	(0.68)		(0.8)	(0.29)		(0.37)
λ (inverse Mills ratio)			−311.73*			−119.92*			−134.34*
			(16.6)			(2.17)			(1.15)
Adj. R-square	0.8086		0.6109	0.8399		0.6955	0.871		0.7201
Log likelihood		12769.7			65675.9			320837	

* Indicates significance at the 0.05 level and the numbers in parentheses are standard errors

region increases the competition for market share, and input-market control is likely to enhance the competitive position of firms.

5.2 Results for the proportional hazard model (AFT)

The results from the AFT model are presented in Table 7. The results show that county employment rate and the presence of interstate highways reduced the probability of survival of the firms in all three industries. While the presence of an interstate highway is considered to be an important location factor, it also increases the competition among firms by increasing market accessibility. Similarly, county employment growth indicates that, apart from the growth of existing firms, more firms are entering the market. This also creates more competition, where higher firm entry is associated with higher exit rates.

Lagged employment had a positive and significant impact on firm survival in the goods- and service-producing industries. Firm size also appears to be important for survival in goods- and service-producing industries. This reinforces the argument that larger firms are less likely to fail. There is a strong empirical literature to support the link between the survival of the firms and start-up size (Agarwal and Audretsch 2001; Agarwal and Gort 2002; Audretsch 1995), while other research has shown that current size is important (Disney et al. 2003; Esteve Pérez et al. 2004).

Quality of the labor force had a positive and significant impact on survival in goods- and service-producing industries. The relatively higher employment of professional and technical workers suggests

Table 7 Estimated coefficients for the hazard model (Weibull distribution)

	IT-producing industry	Goods-producing industry	Service-producing industry
Intercept	2.029*	1.920*	1.932*
	(0.054)	(0.014)	(0.008)
Lag employment	0.000	0.001*	0.001*
	(0.001)	(0.000)	(0.000)
Population density	0.000	0.000*	0.000*
	(0.000)	(0.000)	(0.000)
County employment growth	−0.005*	−0.005*	−0.002*
	(0.001)	(0.000)	(0.000)
Highway	−0.027*	−0.011*	−0.014*
	(0.008)	(0.003)	(0.002)
Quality of labor force	0.001	0.001*	0.001*
	(0.001)	(0.000)	(0.000)
Labor intensity	0.004	−0.000	0.042*
	(0.005)	(0.000)	(0.018)
Industry cluster	0.010	0.005*	−0.005*
	(0.01)	(0.002)	(0.003)
Vertical integration	−0.029	0.174*	0.363*
	(0.054)	(0.041)	(0.002)
λ (inverse Mills ratio)	−0.893	0.077	0.067
	(0.602)	(0.112)	(0.064)
Scale	0.103	0.117	0.125
Log likelihood	1461	5661	23843

* Indicates significance at the 0.05 level and the numbers in parentheses are standard errors

the growth of firms and may be a sign of embracing technology, which is likely to increase the competitiveness of the firm. Vertical integration had a positive and significant impact on survival in the goods- and service-producing industries. Backward and forward inter-industry linkages appear crucial for survival of goods- and service-producing industry firms.

Interestingly, the Mills ratio was insignificant in all three industries. Since the AFT model focused on survival duration, the missing data (those firms that did not survive) did not have a significant impact on the survival duration of other firms.

5.3 Comparison of results from selection bias correction and AFT models

In this section, we consider the results from the selection bias correction and AFT models to determine the most appropriate approach for survival analysis. Both models were used to identify sample selection bias in firm-survival analysis in the case where firms that cease to exist are excluded from the data. Among these two approaches, the Cox AFT model is preferred in firm survival analysis for three reasons.

The first reason for preferring the Cox AFT approach is due to the level of significance of the inverse Mills ratio coefficient. Since the coefficient was significant in the Heckman two-step model, we rejected the null hypothesis (no selection bias) and used the Heckman two-step method to correct the selection bias problem. The results from Cox AFT model suggested that sample selection bias was not a problem in the estimation procedure. In the Cox AFT model, we include all the data in the analysis focusing on survival duration of all of the firms.

Second, the Heckman two-step estimation is still an OLS model, where the error term is assumed to be Independently and Identically Distributed (IID). Generally, such models are not tested for IID assumption validity. However, in the AFT model, the baseline hazard $[h_0(t)]$ is unspecified, which reduces the potential error of incorrect specification (Cleves et al. 2004). Finally, we prefer the AFT approach because the estimated coefficients are more consistent in terms of the direction of the impact on firm survival in all three of the industries studied, while the coefficients were relatively more inconsistent in Heckman two-step model.

6 Conclusion

In this paper, the rates of new firm death and the factors that influence survival were systematically examined using OLS and proportional hazard models (AFT). Technology-intensive firms were more likely to fail within the first 5 years of entry. The two-step OLS results indicated that omitted observations or missing data not only lead to the loss of valuable information, but the estimates derived without that information may lead to erroneous conclusions. Estimation of the inverse Mill's ratio and use of the variable in a firm employment growth model showed that omitted observations can lead to biased coefficient estimates and considerably inflate the model *r*-square values. These results were consistent across all of the models estimated using the Heckman two-step OLS procedures.

On the other hand, the AFT model results suggest that the explanatory variables (from the two-step OLS procedures) had a varying impact on the survival duration of the firm more so than on the growth of the firm. Consistent with other research, the previous level of employment is not only important for firm growth, but for the survival duration of the firm as well.

References

Acs, Z. J. (2002). *Innovations and the growth of cities*. Northampton, MA: Edward Elgar.

Acs, Z. J., Armington, C., & Zhang, T. (2007). The determinants of new-firm survival across regional economies. *Annals of Regional Science, 86*(3), 367–391.

Agarwal, R., & Audretsch, D. B. (2001). Does entry size matter? The impact of the life cycle and technology on firm survival. *Journal of Industrial Economics, 49*, 21–43.

Agarwal, R., & Gort, M. (2002). Technological change. Firm and product life cycle and firm survival. *American Economics Review, 92*, 184–190.

Anagnostaki, V., & Louri, H. (1995). Manufacturing entry in Greece, 1982–1988: Did sectoral policy work? *Journal of Economic Studies, 22*(6), 60–68.

Arauzo, J. M., Manjón, M., Martín, M., & Segarra, A. (2007). Regional and sector specific determinants of industry dynamics and the displacement–replacement effects. *Empirica, 34*, 89–115.

Audretsch, D. B. (1991). New-firm survival and the technological regime. *The Review of Economics and Statistics, 73*, 441–450.

Audretsch, D. B. (1995). *Innovation and industry evolution*. Cambridge, MA: The MIT Press.

Audretsch, D. B., Carree, M. A. Van Stel, A. J. & Thurik, A. R. (2001). *Impeded industrial restructuring: The growth penalty*. Institute for Development Strategies, Indiana University, 1315 East Tenth Street, Room 201, Bloomington, Indiana.

Audretsch, D. B., & Mata, J. (1995). The post-entry performance of firms: Introduction. *International Journal of Industrial Organization, 13*(4), 413–419.

Audretsch, D. B., & Thurik, R. (Eds.). (1999). *Innovation, industry evolution, and employment*. Cambridge: Cambridge University Press.

Baldwin, J. R. (1995). *The dynamics of industrial competition*. Cambridge: Cambridge University Press.

Beck, N. (1992). *Shifting gears: Thriving in the new economy*. Toronto: Harper.

Chakravorty, S. (2003). Capital source and the location of industrial investment: A tale of divergence from post-reform India. *Journal of International Development, 15*, 365–383.

Churchill, N. & Lewis, V. (1983). The five stages of small business growth. *Harvard Business Review, 61* (May-June), 30-50.

Cleves, M., Gould, W., & Gutierrez, R. (2004). *An introduction to survival analysis using stata* (rev. ed ed.). College Station, TX: Stata Press.

Costa, M. T., Segarra, A., & Viladecans, E. (2004). The location of new firms and the life cycle of industries. *Small Business Economics, 22*, 265–281.

Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B34*, 187–220.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm (with discussion). *Journal of Royal Statistical Society, 39*, 1–38.

Disney, R., Haskel, J., & Heden, Y. (2003). Entry, exit and establishment survival in UK manufacturing. *Journal of Industrial Economics, 51*, 93–115.

Dobrev, S. (2001). Revisiting organizational legitimation: Cognitive diffusion and sociopolitical factors in the evolution of Bulgarian newspaper enterprises, 1846–1992. *Organization Studies, 22*, 419–444.

Doi, N. (1999). The determinants of firm exit in Japanese manufacturing industries. *Small Business Economics, 13*, 331–337.

Dunne, T., & Roberts, M. J. (1991). Variation in producer turnover across US manufacturing industries in entry and market contestability. In A. Geroski & J. Schwalbach (Eds.), *Entry and market contestability*. Oxford: Blackwell.

Esteve Pérez, S., Sanchis LLopis, A., & Sanchos LLops, J. A. (2004). The determinants of survival of Spanish manufacturing firms. *Review of Industrial Organisation, 25*(3), 251–273.

Geroski, P. (1991). *Market dynamics and entry*. Oxford: Blackwell.
Grossman, S. J., & Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy, 94*, 691–719.
Hanks, S., Watson, C., Jansen, E. & Chandler, G. (1993). Tightening the life cycle construct: A taxonomic study of growth stage configurations in high technology organizations. *Entrepreneurship Theory & Practice*, *Winter*, 5-29.
Headd, B. (2003). Redefining business success: Distinguishing between closure and failure. *Small Business Economics, 21*, 51–61.
Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator of such models. *Annals of Economic and Social Measurement, 5*, 475–492.
Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.
Heckman, J. (1990). Varieties of selection bias. *American Economic Review, 80*(2), 313–318.
Henry, D., Cooke, S., Buckley, P., Dumagan, J., Gill, G., Pastore, D., et al. (1999). *The emerging digital economy II*. Washington, DC: US Department of Commerce. Economics and Statistics Administration, Office of Policy Development.
Jain, B., & Kini, O. (2000). Does the presence of venture capitalists improve the survival profile of IPO firms? *Journal of Business Finance and Accounting, 27*, 1139–1176.
Keeble, D., & Walker, S. (1994). New firms, small firms and dead firms: Spatial patterns and determinants in the United Kingdom. *Regional Studies, 28*(4), 411–427.
Kimberly, J., & Miles, R. (1980). *The organizational life cycle*. San Francisco, CA: Jossey-Bass Publishers.
Kleijweg, J. M. A., & Lever, M. H. C. (1996). Entry and exit in Dutch manufacturing industries. *Review of Industrial Organization, 11*(3), 375–382.
Knaup, A. E. & Piazza, M. C. (2007). Business Employment Dynamics data: Survival and longevity, II. *Monthly Labor Review,* (September), 3–10.
Lay, T.-J. (2003). The determinants of and interaction between entry and exit in Taiwan's manufacturing. *Small Business Economics, 20*(4), 319–334.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.
Marcus, M. (1967). Firms' exit rates and their determinants. *Journal of Industrial Economics, 16*, 10–22.
MIG, Inc. (1999). *IMPLAN professional, version 2.0: User's guide, analysis guide, data guide*. Stillwater, MN: Minnesota IMPLAN Group, Inc.
Mintzberg, H. (1979). *The structuring of organizations*. Englewood Cliffs, NJ: Prentice-Hall.
Mintzberg, H. (1989). *Mintzberg on management*. New York: The Free Press.
Mitchell, W. (1994). The dynamics of evolving markets: The effect of business sales and age on dissolutions and divestitures. *Administrative Science Quarterly, 39*, 575–602.
Office of Management, Budget. (2000). Standards for defining metropolitan and micropolitan statistical areas; notice. *Federal Register, 65*(249), 82228–82238.
Orr, D. (1974). The determinants of entry: A study of the Canadian manufacturing industries. *Review of Economics and Statistics, 56*(1), 58–66.
Persson, H. (2004). The survival and growth of new establishments in Sweden 1987–1995. *Small Business Economics, 23*(5), 423–440.
Peteraf, M., & Shanley, M. (1997). Getting to know you: A theory of strategic group identity. *Strategic Management Journal, 18*, 165–186.
Praag, C. M. V. (2003). Business survival and success of young small business owners. *Small Business Economics, 21*, 1–17.
Quinn, R., & Cameron, K. (1983). Organizational life cycles and shifting criteria of effectiveness: Some preliminary evidence. *Management Science, 29*, 33–41.
Reid, G. (1995). Early life-cycle behaviour of micro-firms in Scotland. *Small Business Economics, 7*(2), 89–95.
Reuber, A., & Fischer, E. (1999). Understanding the consequences of founders' experience. *Journal of Small Business Management, 37*(2), 30–45.
Reynolds, P., Storey, D. J., & Westhead, P. (1994). Cross-national comparisons of the variation in new firm formation rates. *Regional Studies, 28*(4), 443–456.
Santarelli, E., & D'Altri, S. (2003). The diffusion of e-commerce among SMEs: Theoretical implications and empirical evidence. *Small Business Economics, 21*(3), 273–283.
Shaffer, R., Deller, S. C., & Marcouiller, D. W. (2005). *Community economics: Economic structure and change in smaller communities*. Ames, IA: Iowa State University Press.
Shapiro, D., & Khemani, R. S. (1987). The determinants of entry and exit reconsidered. *International Journal of Industrial Organization, 5*, 15–26.
Storey, D. J. (2000). *Small business critical perspectives on business and management*. New York: Routledge.
Strotmann, H. (2007). Entrepreneurial survival. *Small Business Economics, 28*, 87–104.
Thompson, P. (2005). Selection and firm survival: Evidence from the shipbuilding industry, 1825–1914. *Review of Economics and Statistics, 87*, 26–36.
Tichy, N. (1980). Problem cycles in organizations and the management of change. In J. Kimberly & R. Miles (Eds.), *The organizational life cycle*. San Francisco, CA: Jossey-Bass Publishers.
US Small Business Administration. (2008). *Table of small business size standards matched to the North American industry classification system codes*. Washington, DC: US Small Business Administration.
Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources, 33*(1), 127–169.
Walker, G., & Weber, D. (1984). A transaction cost approach to make-or-buy decisions. *Administrative Science Quarterly, 2*, 373–391.
Walker, G., & Weber, D. (1987). Supplier competition, uncertainty and make-or-buy decisions. *Academy of Management Journal, 30*, 589–596.
Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*(1), 327–350.
Woods and Poole. (2006). *Complete economic and demographic data source*. Washington, DC: Woods and Poole, Inc.