A New Index of Entrepreneurship Measure

Annamaria Bianchi
University of Bergamo

Silvia Biffignandi
University of Bergamo

A new index of entrepreneurship measure based on M-quantile regression methods is introduced. This index allows meaningful comparison of the entrepreneurial performances. The new method is illustrated using the Kauffman Firm Survey, a study of new businesses in the United States. A comparative analysis between this index and a classical index obtained with Data Envelopment Analysis is performed. Using this index, possible determinants of entrepreneurship are investigated applying beta regression.

INTRODUCTION

Entrepreneurship is recognized as a critical factor for the wealth and competitiveness of a nation. It is a focal point of public policy and policy-makers have made promoting entrepreneurship a priority. Special attention is focused on how government policies and other factors can influence the amount and type of entrepreneurship. The interest in this problem is also confirmed by the attention that many international bodies (like OECD - the Organization for Economic Co-operation and Development - and Eurostat) are devoting to the study of this phenomenon (OECD/Eurostat, 2009).

Despite its acknowledged importance in the economic process, entrepreneurship is an elusive phenomenon, difficult to define and measure. Existing research is far from conclusive in terms of providing a comprehensive definition and measure of entrepreneurship. However, if entrepreneurship is not properly measured, then it is not possible to understand its determinants, why and where the phenomenon flourishes, and to assess the effects of public policies.

Being able to measure entrepreneurship is important both as a competitive factor for companies and in planning development policies. On the one hand, it is useful for searching good entrepreneurs as it identifies factors on which to focus to make performing companies. On the other hand, it is useful for policy makers to identify the priority areas of intervention.

In this paper we focus on the problem of measuring entrepreneurship at the firm-level. We propose a new quantitative index of entrepreneurship measure based on M-quantile regression. It is a production function approach that relies on the idea that a firm transforms inputs into outputs and the more outputs a business can produce with fewer inputs the more successful the entrepreneur is. The idea behind this entrepreneurship indicator is that the performance of the entrepreneur is strictly connected to the performance of the business. The index is then compared to a common indicator currently used and obtained from the application of Data Envelopment Analysis (DEA) techniques. We show that the new index allows overcoming many drawbacks presented by conventional measures.

The empirical analysis is based on the Kauffman Firm Survey (KFS) data. This survey is sponsored by the Ewing Marion Kauffman Foundation. The KFS is the largest longitudinal survey of new businesses in the world. It consists of a cohort of 4,928 firms that were started in 2004 in the United States and tracked over the following years. This dataset presents a number of advantages for our study, since it contains detailed information on the business, the owner's personal characteristics and the financial data. The KFS focus on businesses in their early years of operation. We believe that this dataset is especially suited for studying entrepreneurship. Indeed, entrepreneurship is especially critical at the beginning of a business. As it is well know, birth and death of enterprises are quite high and critical survival time is from three to five years. Thus the performance of the new businesses in the short time (three to five years) is a crucial indicator of entrepreneurship.

The rest of the paper is organized as follows. In the next section we briefly review some of the different entrepreneurship measures that have been proposed in the literature for empirical firm-level studies. Next we propose a new method for the evaluation of entrepreneurship. The methodology is then applied to the Kauffman Firm Survey database and compared to an index obtained with DEA techniques. A second stage regression is also considered to measure the impact of different factors on the measured entrepreneurship.

BACKGROUND LITERATURE

Literature is reach on empirical studies aimed at measuring entrepreneurship and determining its factors. Various levels of measurement have been proposed. In this section we briefly review the principal ones. For more details on the historical development of the concept and measurement of entrepreneurship, please refer to Foreman-Peck (2005).

One approach uses a nominal scale and measures entrepreneurship using binary variables of success/no success or self-employment/employment. Next probit models are run to study the relationship between a number of possible determinants (human capital, initial business size, access to finance) and the defined performance (Evans & Leighton, 1989, Cressy, 1996, Blanchflower & Oswald, 1998, and Harada, 2003). Although these performance measures have the advantage of being readily available in surveys, they are not necessarily good measures of entrepreneurship. Indeed, a measure of entrepreneurship requires for a link with effective innovation and it is clear that not all start-ups involve innovation.

A higher level of measurement is represented by the ordinal scale. Entrepreneurship measures based on ordinal scale are necessary to be able to answer questions like 'How well do entrepreneurs do what they do?'. These kinds of measures allow to assess whether an entrepreneur is more or less successful than another one. A multinomial logit is then used to identify determinants (Grilo & Thurik, 2008).

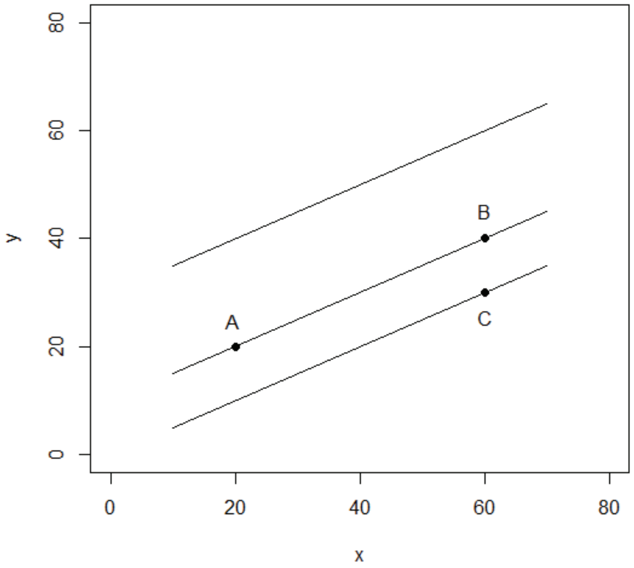
A more demanding level of measurement is the ratio scale, which allows comparisons of the amounts of entrepreneurship. In this category profit, income or wealth generated by the business are contemplated as possible indicators of entrepreneurial performance in particular contexts. These kinds of measure are subject to a number of flaws, mainly connected to the problem of measuring the wealth of an entrepreneur.

More sophisticated methods are based on production functions and, more specifically, on the relationship between observed production and some ideal or potential production, represented by the so-called production frontier. Suppose that the value of the output from a certain business can be measured by a variable y and associated with this is a vector x of inputs values, with the relationship between y and x determined by an appropriately chosen production function. Production frontier corresponds to best practice and it is defined as the maximum achievable output $f(x)$ from input x. Adopting an output-oriented approach, entrepreneurship measures are then defined comparing the business's total value of output to this frontier. Usually the comparison takes the form of the ratio of observed to maximum potential output obtainable from the given input, that is, $y / f(x)$. Different approaches are classified according to the method used for the derivation of the production frontier. The main distinction is between stochastic and deterministic methods. Methods based on stochastic frontier production functions, which were introduced independently by Aigner et al. (1977) and Meeusen & van den Broeck (1977), belong to the first group. They involve the estimation of a stochastic production function, where the output of a firm is a function of a set of inputs, inefficiency and random error. This method has the advantage of allowing statistical inference, but it has the disadvantage of strong parametric assumptions both on the functional form of the frontier and the distribution of the data. In this context various methods have been proposed for the estimation of f, but all of them may lack robustness if the assumptions are not satisfied. In particular, outliers in the data may unduly affect the estimate of f.

Alternative deterministic approaches, which overcome some of these problems, are the linear programming techniques of Data Envelopment Analysis (DEA) (Charnes et al., 1978) and Free Disposal Hull (FDH) (Deprins et al., 1984). These are nonparametric approaches: they do not impose any assumptions about the functional form hence they are less prone to miss-specification. These kinds of measures are widely used as entrepreneurship indicators. See, for example, Foreman-Peck (2005). However, because the DEA and FDH estimates of f depend critically on the extreme observations, these techniques are prone to the influence of outliers. Moreover an important flaw of these approaches is that they are seen as non-statistical, not distinguishing inefficiency from random shocks. Efforts to add a stochastic dimension to these methods have been made along several lines. For a review of these extensions, please refer to Ray (2004).

The most important drawbacks of benchmarking measures (both stochastic and deterministic) are that (i) they are not robust to outliers and (ii) they can result in quite different values for two businesses operating at the same level of efficiency but with different levels of inputs. Concerning the problem of outliers, methods have been developed for their identification in this context. See, for example Wilson (1993). However, the methods are non-robust in nature. Concerning the second point, consider the following example that was proposed by Kokic et al. (1997) to explain this problem in the single-input, single-output case.

FIGURE 1
INPUT-OUTPUT RELATION



Note. The lower, central and upper lines correspond to inefficiency frontier, average production function, and the efficiency frontier, respectively.

Figure 1 represents a possible input-output relation. For simplicity, we consider the homoschedastic case. The lower and upper lines correspond to production inefficiency and efficiency, respectively. Businesses that lie on the bottom line are operating at their most inefficient level, while businesses lying on the upper line are operating at their most efficient level (the equation for this line is $f(x)$). The central line represents the 'average' production frontier, i.e. the 'average' entrepreneurial level.

Since businesses A and B lie on this line, they should have the same entrepreneurship index. The entrepreneurship indexes associated to A and B by benchmarking measures, however, are 1/2 and 2/3, respectively. Moreover, business C, which lies on the inefficiency frontier, has the same value as business A. It is clear from the example that indexes based on benchmarking measures depend on x. This is a strong limitation of these indexes. Measures of entrepreneurship should not depend on the level of inputs of the business, otherwise they would be biased and hide the true relationship between returns and the scale of business operations. The idea underlying this consideration is that an entrepreneur may be performing well even if he is running a small firm.

PERFORMANCE EVALUATION

In this section we face the problem of introducing an index of entrepreneurship which overcomes the problems presented by many other performance measures currently used. Just like a production-performance measure (Kokic et al., 1997), we argue that an entrepreneurship index should satisfy the following four properties:

1. It must lie between 0 and 1.
2. The poorest performing entrepreneurs should have indexes close to 0.
3. The best performing entrepreneurs should have indexes close to 1.
4. The distribution of the entrepreneurship indexes should not depend on the level of inputs x of the business.

As shown in the previous section, the last property is not satisfied in general by indexes of the form $y / f(x)$, resulting in quite different values for two entrepreneurs operating at the same level of efficiency but with different levels of inputs. To say it differently, the values of benchmarking indexes generally increase with the level of inputs of the business run by the entrepreneur. Since the scale of a business operation is usually related to the amount of inputs, if an entrepreneurship index does not satisfy criterion 4, then scale effects are convoluted with input value effects. However, not necessarily an entrepreneur is performing well just because he is running a large firm. Methods that fail criterion 4 are therefore not appropriate for investigating returns to scale - that is, the effect of business size on entrepreneurship. For this reason, they are not reliable and they cannot be used for investigating how scale may affect the entrepreneurial performance. In the example presented in the previous section, benchmarking measures are positively correlated to business size even though there are no increasing returns to scale.

A measure of production performance with properties 1-4 can be obtained from the application of M-quantile regression. We decided to adopt this method to construct an index of entrepreneurship measure. This is an innovative approach to the measurement of entrepreneurship. Indeed to the best of our knowledge there is no application of such method in this context.

This index presents several appealing characteristics. It is outlier robust, it does not depend on the level of inputs of the business, and it has a stochastic nature by construction.

Before giving the exact definition of the index, we briefly introduce M-quantile regression basic concepts. M-quantile regression was proposed by Breckling & Chambers (1998) and it is a generalization of M-regression, which was introduced by Huber (1973) in answer to the excessive sensitivity of ordinary least square (OLS) regression to small deviations from the model assumptions. Consider a dependent variable y taking real values and a set of k-dimensional explanatory variables x and suppose to observe the values (x_i, y_i), $i = 1, \ldots, n$. Assuming a linear model for the data, the M-regression estimates are defined as the solution $\hat{\beta}$ of the estimating equation

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i^T \beta}{s}\right) x_i = 0, \tag{1}$$

where ψ is called 'influence function' and s is a suitable robust estimator of scale, such as the median absolute deviation estimator to be defined later. Throughout this article we use the Huber Proposal 2 influence function

$$\psi(u) = \begin{cases} u & if \ |u| \le c \\ c\,\mathrm{sgn}(u) & if \ |u| > c \end{cases} \tag{2}$$

where c is a properly chosen positive constant. Other specifications for the function ψ are possible. Notice that M-regression includes as special cases median regression ($c = 0$) and OLS regression ($c \to +\infty$).

Similarly, M-quantile regression is defined as the solution $\hat{\beta}_q$ of (1) with the influence function given by

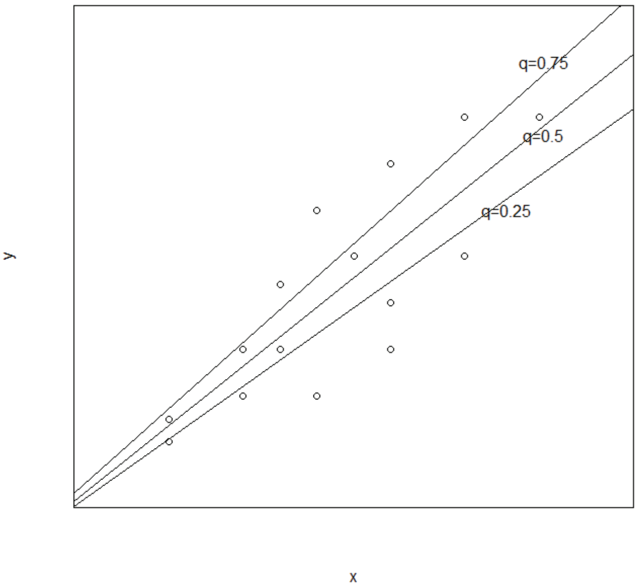
$$\psi_q(u) = \begin{cases} q\psi(u) & u \ge 0 \\ (1-q)\psi(u) & otherwise \end{cases} \tag{3}$$

Special cases include quantile regression for $c = 0$ (Koenker & Bassett, 1978) and expectile regression for $c \to +\infty$ (Newey & Powell, 1987). However, M-quantile regression is preferable to both quantile and expectile regression. Indeed, unlike expectile regression, it is not unduly influenced by extremely large residuals; moreover, it overcomes the problems of quantile regression associated with the use of a discontinuous influence function.

The idea on which this method is based is that the M-quantile regression hyperplane $x^T \hat{\beta}_q$ corresponding to a particular value of q is obtained by weighting positive residuals by a factor q and negative residuals by a factor $(1-q)$. For example, when $q = 0.25$ negative residuals have three times the weight of positive residuals in the estimating equation for the corresponding M-quantile. As a result this plane will lie below the ordinary M-regression hyperplane. In the single-input, single-output case, Figure 2 shows the M-quantiles fitted on a set of input-output combinations at 0.25, 0.5 and 0.75.

More generally, M-quantile regression leads to a family of hyperplanes indexed by the value of the corresponding quantile coefficient q. For each value of q in $(0,1)$, the corresponding model $x^T \hat{\beta}_q$ shows how the M-quantile of order q of the conditional distribution of y given x varies with x. For large values of q the M-quantile surface describes the 'average' output of relatively efficiently performing businesses, and for small q it describes the 'average' output of relatively inefficiently performing businesses. For a certain business with input x_i and output y_i, it is therefore natural to define an entrepreneurship index as the value q of the M-quantile surface passing through the observation (x_i, y_i).

FIGURE 2
FITTED M-QUANTILES AT 0.25, 0.5 AND 0.75.



Taking the abovementioned characteristics into account, we define the Index of Entrepreneurship Measure (*IEM*, hereafter) for unit i as $IEM_i = q$ if $y_i = x_i^T \hat{\beta}_q$.

It is clear that the definition of the index rests on the assumption that the underlying production model is approximately true for the data. This assumption may be checked by simple graphical techniques.

In practice, a common problem raising at this point is that the fitted hyperplanes so obtained can cross over for different values of q. In this case it is clear that the use of M-quantile surfaces to define the index turns out to be problematic. Crossing of M-quantile surfaces reflects the paucity of data in the region concerned. It is a finite-sample problem and it is typically due to a combination of variability of the estimates, regression model misspecification and collinearity in the explicative variables. If crossing occurs we define the *IEM* index following the method proposed by Kokic et al. (1997). Denote S_i the set of those values of q such that the q-th M-quantile passes through y_i. The index of entrepreneurship measure is then defined as the value in S_i closest to 0.5. Notice that this is a conservative approach since we define the index as the value closest to the "mean". Another possible approach to the problem of M-quantile crossing consists in ensuring monotone fitted M-quantile regression surfaces (He, 1997).

Notice that by construction this index has a stochastic nature, thus allowing statistical inference on the computed values. It is straightforward to prove that the asymptotic distribution of the index is uniform in the case of quantile regression ($c = 0$). In the more general case of M-quantile regression, empirical evidence shows that the asymptotic distribution resembles a Beta random variable. The proof of the asymptotic properties of the index goes beyond the scope of this work and it will be the subject of further research.

EMPIRICAL ANALYSIS

In the present section first we introduce the dataset used for the empirical analysis, the Kauffman Firm Survey (KFS). Next we compute the *IEM* index of entrepreneurship measure for firms in the KFS database. We analyze its characteristics and compare them with those of the classical entrepreneurship index obtained with DEA methods. Finally, we study the impact of various factors on the index by means of a beta regression.

The Kauffman Firm Survey

The Kauffman Firm Survey (KFS) is a longitudinal study of new businesses in the United States. A new business is defined as a new, independent business that was created by a single person or a team of people, the purchase of an existing business or the purchase of a franchise. Businesses that were inherited from someone else, were set up as a subsidiary of an existing business, or created as a not-for-profit organization were excluded. At present six years of data are available, from 2004 (foundation year) until 2009. Additional years are planned. The empirical analysis is performed on the 2004 data.

To create the panel a random sample was drawn from the Dun & Bradstreet database. In response to the Kauffman Foundation's interest in understanding the dynamics of high-technology businesses, the KFS oversampled these businesses. The Baseline Survey was conducted between July 2005-July 2006 with data collection reference year 2004. As already said, it contains information on 4,928 firms. For more details, see Robb et al. (2010).

The survey database includes detailed information both on the firm and up to ten business owners per firm. We use the KFS restricted-access data file which contains more detailed information, including industry and geographical coding, as well as other variables. Detailed information on the firm includes industry, physical location, employment, profits, intellectual property and financial capital. Information on the owners includes age, gender, race, ethnicity, education, work experience and previous start up experience. One potential issue for the restricted-access data file is that firms can choose to disclose financial information using range values. For example, firms can only provide range values for total cost and total revenue, while other firms provide exact values for the same variables. In order to make the KFS dataset more complete, we transform these range values into numeric values by using the central point for each range class.

Since the KFS provides the socio-demographic information of up to ten owners, to assign entrepreneur demographics at the firm level we defined a primary owner who presumably has the most influential managerial power over the business. For firms with multiple owners, the primary owner is the one that has the largest share of the business. In cases where two or more owners owned equal shares, hours worked and a series of other variables are used to create a rank order of owners in order to define a primary owner. Once the primary owner is identified, her/his characteristics are assigned to the firm. They include gender, race, ethnicity, age, education and entrepreneurial activities before she/he started the business in 2004. As prior entrepreneurial activities, we consider both start-up experience and years of prior work experience in the same industry as the current business. The list of variables together with their description is given subsequently (Table 2).

IEM Index

The *IEM* index is based on the identification of an approximation to the true production function model relating the inputs and the output of the business. The model used in the analysis is the Cobb-Douglas type production function

$$\log(REV_i) = \beta_0 + \beta_1 \log(EXP_i) + \beta_2 \log(HOURS_i), \tag{4}$$

where REV is the total revenue of the business, EXP is total cost and $HOURS$ represents owner hours. The total revenue variable measures the output of the business. Total cost is a measure of the resources used by the enterprise, while the variable owner hours measures the effort of the entrepreneur(s). The variable $HOURS$ is the sum over all owners.

Starting from the production function model (4) we estimated M-quantile regression planes over a grid of values of q, specifically for $q \in \{0.02, 0.04, \dots, 0.96, 0.98\}$:

$$\mu_q(EXP, HOURS) = \beta_{0q} + \beta_{1q}\log(EXP_i) + \beta_{2q}\log(HOURS_i),$$

where $\mu_q(EXP, HOURS)$ is the hyperplane corresponding to the q-th regression M-quantile plane of the log of total revenue. To obtain the regression fits we used a modification of the *rlm* function in R (Venables & Ripley, 2002). We used the KFS data for 2004. Some businesses were excluded from the estimation procedure because some of the variables were not available. Adjusted survey weights were used in the estimation to take into account the stratified sample design used in the KFS and the location and response adjustments. The influence function used was Huber's Proposal 2 (defined by (2) and (3)) with $c = 1.345$ and s given by the median absolute deviation (MAD) of the residuals

$$s = \frac{MED\left\{|\,(y_i - x_i^T\hat{\beta}_q) - MED\,|\,y_i - x_i^T\hat{\beta}_q\,||; i = 1,\ldots,n\right\}}{0.6745}.$$

Since the M-quantile surfaces crossed over for different values of q, we applied the method described in the previous section by means of the following algorithm. If the value $\log y_i$ was below the 0.5 M-quantile plane, then we searched for the value q closest to 0.5 in {0.02,0.04,....,0.48} such that $\log y_i$ was above the q-th M-quantile plane. If such a q existed, we defined $IEM_i = q + 0.01$, else $IEM_i = 0.01$. Similarly, when $\log y_i$ was on or above the 0.5 M-quantile plane, we found the value q closest to 0.5 in {0.50,0.52,...0.98} such that $\log y_i$ was below the q-th M-quantile plane. If it existed, we defined $IEM_i = q - 0.01$, otherwise $IEM_i = 0.99$.

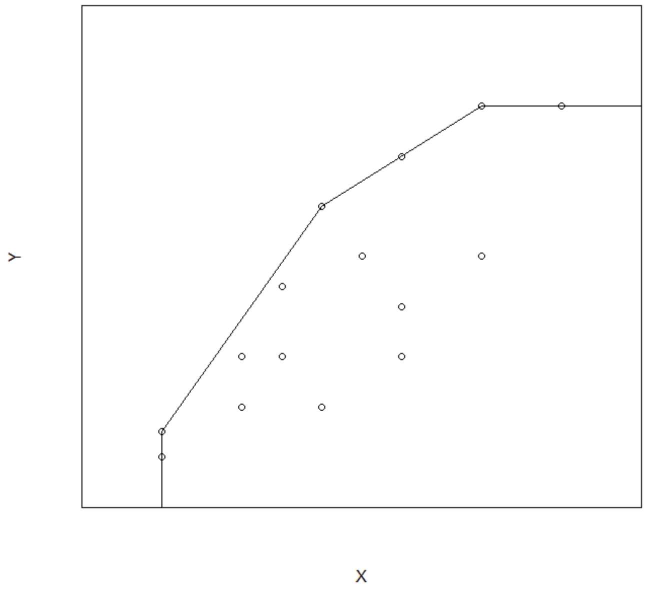
The index was computed for 2975 firms. Next, we checked criterion 4 by examining scatter-plots of the index *IEM* against the variables *EXP* and *HOURS*. These scatter-plots show that there is no relationship between the index and the variables. Unfortunately they cannot be shown due to confidentiality of the data.

In order to analyze the difference of the proposed methodology with benchmarking indexes, we performed an empirical comparative analysis of the *IEM* index with an index based on Data Envelopment Analysis (DEA). DEA combines the estimation of an efficient frontier with the measurement of performance related to this frontier. With knowledge of all businesses' input-output combinations, it is possible to compute which entrepreneurs are efficient (i.e. their business is on a frontier) and which are inefficient or dominated. An efficiency benchmarking requires not only information about inputs and outputs, but also weights and formulae for combining them. The basic DEA models mainly differ in the assumption that they make about the efficient frontier and the production possibility set -- that is, the set of all feasible input-output combinations. By way of example, we adopt the method based on DEA with variable returns to scale (VRS) which assumes free disposability of inputs and outputs, convexity and variable returns to scale. In this context, an entrepreneur is efficient if no linear combination of other entrepreneurs gives a higher efficiency score for their business. The 'efficient frontier' consists therefore of linear combinations of efficient entrepreneurs. Linear programming is needed to identify the above mentioned efficient frontier (Bogetoft & Otto, 2011).

In the case of single-input, single-output, Figure 3 shows the efficient frontier generated by a set of businesses' input-output combinations in case of VRS.

Once the efficient frontier has been determined, DEA measures the entrepreneurial performance relatively to this frontier. In the sequel, we adopt an output-oriented approach, that is, we assume that the objective of an entrepreneur is to produce the maximum quantity of output from a specified input bundle. The DEA index is then obtained by comparing the actual output produced with the corresponding benchmarking quantity.

FIGURE 3
DEA EFFICIENT FRONTIER UNDER VARIABLE RETURNS TO SCALE



We compute the DEA indexes of entrepreneurship for the KFS data by means of the Benchmarking package in R (Bogetoft & Otto, 2011). In order to obtain comparable results, we consider the same variables used for the computation of the *IEM* index measured in logarithmic units. Before the computation, we checked for the presence of outliers using the method proposed in Wilson (1993) and found that there were any.

Data were analyzed using the SAS System for Windows (release 9.2; SAS Institute Inc, Cary, NC). All analysis included sample weights that account for the unequal probabilities of selection because of stratified sampling and nonresponse.

First we performed a Spearman correlation rank test to analyze concordance of the two indexes in ranking firms. The value for the test statistic was 0.7507, denoting that the indexes partially agree, but still there are differences in ranking. Looking at the single values, we argue that these differences are mainly due to the different behaviour of micro-firms and to the fact that the DEA index tends to assign higher performances to entrepreneurs running large firms. A descriptive analysis of *IEM* and DEA indexes is given in Table 1.

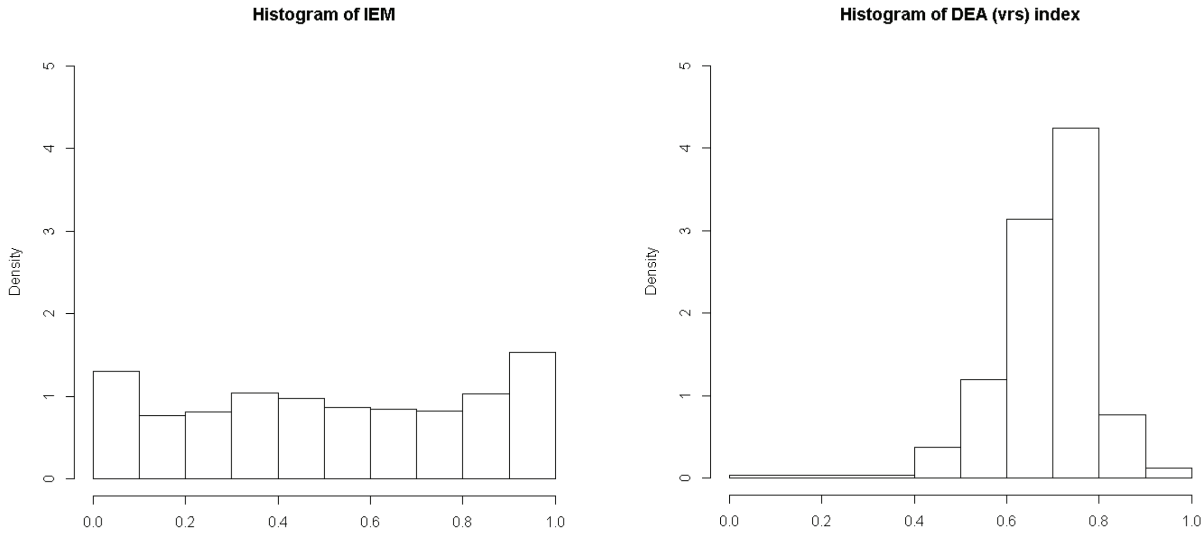
TABLE 1
SUMMARY STATISTICS OF *IEM* AND DEA ENTREPRENEURSHIP INDEXES

Index	Mean	95% CI	Q_1	Q_2	Q_3
IEM	0.515	0.502-0.528	0.253	0.509	0.803
DEA	0.687	0.511-0.517	0.485	0.516	0.545

The mean value of *IEM* is 0.515, while DEA index mean value is 0.687. Looking at the first and third quartiles of the distributions of the indexes, it is clear that *IEM* presents much larger variability. Comparing the means with the corresponding medians of the distributions, we see that DEA is more skewed. These conclusions are more evident looking at the histograms of the distributions (Figure 4). It is also evident that the *IEM* index allows detecting greater differentiation among the performances.

Concerning the *IEM* index, a *q-q* plot suggests that its distribution is consistent with the beta random variable.

FIGURE 4
HISTOGRAMS FOR *IEM* AND DEA INDEXES



In order to analyze the effect of scale on entrepreneurial performance, the indexes were locally regressed against firm size using PROC LOESS in SAS. The smoothing parameter was selected according to the bias corrected Akaike information (Hurvich & Simonoff, 1998). Firm size is measured in terms of the log of the number of employees. Notice that we do not expect many returns to scale, as we are only considering small/medium enterprises (just up to over one hundred and fifty employees). Figure 5 shows the curves obtained for *IEM* index (solid line) and DEA index (dashed line).

FIGURE 5
NONPARAMETRIC REGRESSION OF *IEM* INDEX (SOLID LINE) AND DEA INDEX (DASHED LINE) AGAINST FIRM SIZE

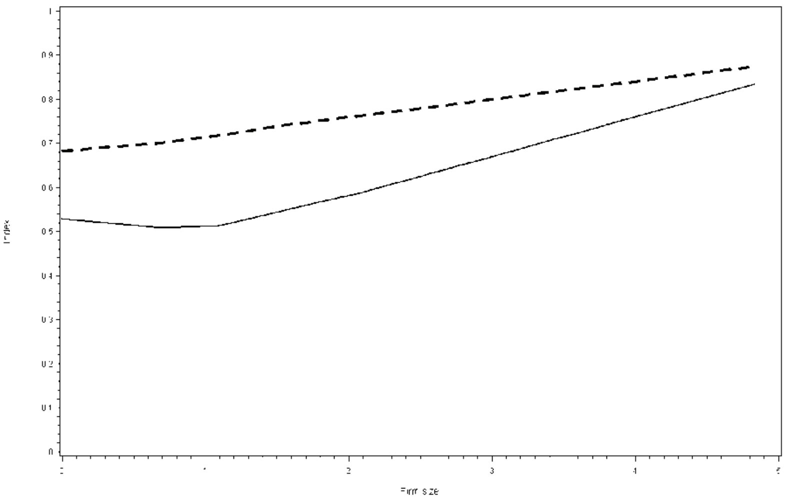


TABLE 2
LIST OF POSSIBLE FACTORS: NAME, DESCRIPTION

Variable	Description and codes
FEM	Gender of the primary owner, 1=female, 0=male
AGE	Age of the primary owner
EDU	Highest level of education completed by the primary owner, measured with a scale ranging from 1 (=Less than 9th grade) to 10 (=Professional school or doctorate)
WORKEXP	Years of work experience in the same business
STARTUP	Number of new businesses started by the primary owner
EMP	Is primary owner also a paid employee at business? 1=Yes, 0=No
HISP	Is primary owner of Hispanic origins? 1=Yes, 0=No
WHITE	Is primary owner white? 1=Yes, 0=No
PAT	Does business have patents? 1=Yes, 0=No
IP	Does business have intellectual properties (patents, copyrights, trademarks)? 1=Yes, 2=No
TOTIP	Total number of intellectual properties the business possesses
HT	High technology industry indicator based on whether the business is a Technology Employer or Technology Generator, 1=High tech, 0=Non-high tech
HOME	Is the business homebased? 1=Yes, 0=No
MULTIOWN	Is the business multiowned? 1=Yes, 0=No
NUMEMP	Number of employees

Considering *IEM* index, we see a differentiation between firms with less than three employees approximately and firms with more than three employees. This is interesting since one to three employees size corresponds to micro-firms, which usually are studied separately because of their particular economic characteristics. Our *IEM* indicator suggests that micro-firms are only sensitive to the performance of the entrepreneur. They are quite insensitive to the number of employees. If one wants to detect an impact, we see that the relationship is slightly decreasing: the higher the number of employees, the lower the possibility of entrepreneurial performance. Then we observe performance slightly increasing with size. The growth is linear. This result can be justified by the fact that, after all, we are considering a well-defined dimension size. On the other hand, DEA index increases with firm size. It is not sensitive to the differences that *IEM* index is able to perceive. *IEM* index grasps immediately the difference between micro-firms and small-medium size firms. It is detecting only limited differences within the small-medium size, i.e. it recognizes the approximately homogeneous entrepreneurial performance in this class size. We expect that, if the dataset contained also data on large firms, this index would be able to perceive differences in entrepreneurs' performances related to larger business size.

Determinants of Entrepreneurship

Finally, as a first step towards understanding how different factors contribute to the variability in the index values (i.e. entrepreneurial performance differences), we modelled the impact of several factors concerning both the entrepreneur and the firm on the index. The list of possible factors that we considered together with their description is given in Table 2.

The summary statistics of these possible factors in the Baseline Survey are presented in Tables 3 and 4. We see that 69.3% entrepreneurs of the study population are men, 56.7% have a college degree or higher level of education, 82.7% are white. The mean age is 44.55, the mean number of other new businesses started by the entrepreneurs is 0.963 and the mean number of years of experience in the same industry of the business is 11.87.

TABLE 3
SUMMARY STATISTICS OF THE ENTREPRENEUR CHARACTERISTICS IN BASELINE SURVEY

	Mean	95% CI
Gender (*FEM*)		
Male	0.693	0.682-0.705
Female	0.307	0.295-0.318
AGE	44.55	44.20-44.90
Education (*EDU*)		
College +	0.567	0.551-0.583
College -	0.433	0.417-0.449
Previous work-experiece (*WORKEXP*)	11.87	11.54-12.20
Previous start-up (*STARTUP*)	0.963	0.882-1.043
Employee owner (*EMP*)		
Paid employee	0.480	0.464-0.497
Not paid employee	0.520	0.503-0.536
Hispanic origin (*HISP*)		
Hispanic origin	0.066	0.058-0.074
Not Hispanic origin	0.934	0.926-0.942
WHITE		
White	0.827	0.815-0.840
Other	0.173	0.160-0.185

As far as the characteristics of the firms are concerned, Table 4 shows that in the study population 5.6% of the businesses are high-tech and 2.2% possess patents. Moreover, 49.2% are homebased businesses and 34.9% are multiowned.

In order to quantify the impact of these factors on entrepreneurship, we performed a beta regression with a logit link specification (Kieschnick & McCullough, 2003). This choice was suggested by the consideration outlined in the previous section that the *IEM* index follows a Beta distribution. Starting from the list of variables given in Table 2, we tested alternative models. The final model includes entrepreneur's characteristics related to gender, education, race, and previous work experience. As far as the business' characteristics are concerned, factors influencing entrepreneurship are patents, high-tech firms, intellectual properties and number of employees. The estimates are given in Table 5.

TABLE 4
SUMMARY STATISTICS OF THE FIRMS CHARACTERISTICS IN BASELINE SURVEY

	Mean	95% CI
Patents (*PAT*)		
Yes	0.022	0.018-0.027
No	0.978	0.973-0.982
Intellectual properties (*IP*)		
Yes	0.192	0.179-0.204
No	0.808	0.796-0.821
Total number of intellectual properties (*TOTIP*)	1.272	0.962-1.582
High-tech (*HT*)		
High-tech	0.056	0.051-0.061
Non-high tech	0.944	0.939-0.949
Homebased (*HOME*)		
Yes	0.492	0.476-0.508
No	0.508	0.492-0.504
Multiowned (*MULTIOWN*)		
Yes	0.349	0.333-0.364
No	0.651	0.636-0.667
Number of employees (*NUMEMP*)	1.869	1.688-2.049

The main findings are the following. Women and non-white entrepreneurs seem to face more difficulties in starting a new business. Education is a positive factor, as well as previous work experience in the same industry. The fact that the entrepreneur is also a paid employee at business has a highly significant positive impact. Entrepreneurs running high-tech firms seem to be favoured. On the other hand, entrepreneurs running multi-owned firms and firms with intellectual properties seem to have worse performances. As far as the variable number of employees is concerned, the results found in the previous section are confirmed.

Since the conditional expectation function is given by a logit, the parameters' estimates are not directly interpretable in terms of marginal effects of a change in a regressor on the entrepreneurship index. The marginal effect of x_j is given by

$$\frac{\partial E(IEM \mid x)}{\partial x_j} = \frac{\hat{\beta}_j \exp\left(-x^T \hat{\beta}\right)}{\left(1+\exp\left(-x^T \hat{\beta}\right)\right)^2}, \quad j = 1, \ldots, k$$

where x represents the vector of regressor variables and $\hat{\beta}$ the corresponding estimated parameters. For example, setting FEM=0, EDU=6 (Bachelor's degree), WHITE=1, IP=0, WORKEXP=11.87 (mean) we can compute the marginal effect of EMP on $E\left(IEM \mid x\right)$ at three different levels for the number of employees: NUMEMP=10, NUMEMP=50, and NUMEMP=150. The estimated marginal effects are 5.9%, 5.4% and 2.3%, respectively.

In future work we would like to extend the *IEM* index to panel data. This would allow us to study the dynamics of the phenomenon, taking into account the heterogeneity between statistical units, and addressing traditional econometric problems in cross-section regression, such as unobserved individual effects, endogeneity, and dynamics.

TABLE 5
RESULTS FOR THE BETA REGRESSION

Variable	Estimate	Std Error	t-value		p-value
INTERCEPT	-0.3415	0.1297	-2.63	**	0.0085
FEM	-0.1721	0.0512	-3.36	***	0.0008
AGE	-0.0012	0.0022	-0.52		0.6009
EDU	0.0236	0.0107	2.21	**	0.0272
WORKEXP	0.0125	0.0023	5.42	***	<0.0001
STARTUP	0.0039	0.0100	0.39		0.6995
EMP	0.2349	0.0440	5.34	***	<0.0001
HISP	0.0507	0.0993	0.51		0.6095
WHITE	0.0329	0.0139	2.37	**	0.0180
PAT	-0.1458	0.1626	-0.90		0.3697
IP	-0.1829	0.0622	-2.94	**	0.0033
TOTIP	-0.0029	0.0027	-1.08		0.2795
HT	0.1366	0.0660	2.07	*	0.0388
HOME	0.0021	0.0447	0.05		0.9618
MULTIOWN	-0.1238	0.0462	-2.68	**	0.0075
NUMEMP	0.0152	0.0035	4.40	***	<0.0001

* p-value<0.1; ** p-value<0.05; *** p-value<0.001

CONCLUSIONS

In this paper, we have proposed an innovative approach for the study of entrepreneurship. A new performance indicator (*IEM*) is introduced in the framework of M-quantile regression. By construction, this index has a stochastic nature and it is outlier robust. Moreover, it does not depend on the level of inputs of the business. An empirical comparison with the standard indicator obtained with DEA techniques is carried out. This comparison showed that DEA index depends on the level of input of the business and therefore it is not possible to use it for studying returns to scale.

As regards substantive results, a beta regression has been performed to analyze how well the entrepreneurs and firm characteristics do explain the observed variation in the entrepreneurial performances. We found that positive factors are education, previous work experience in the same industry, white race and the fact that the entrepreneur is also a paid employee at business.

ACKNOWLEDGMENTS

Paper supported by the ex 60% University of Bergamo, Biffignandi grant and PAADEL project (Lombardy Region – University of Bergamo joint project). The authors are thankful to the Kauffman Foundation for providing training and support. Any opinions, finding, and conclusions or

recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Ewing Marion Kauffman Foundation.

REFERENCES

Aigner, D. J., Lovell, C. A. K. & Schmidt, P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*, 6, 21-37.

Bogetoft, P. & Otto, L. (2011). *Benckmarking with DEA, SFA, and R*, New York: Springer-Verlag.

Blanchflower, D. G. & Oswald, A. J. (1998). What Makes an Entrepreneur. *Journal of Labor Economics*, 16, 26-60.

Breckling, J. & Chambers, R. (1988). M-quantiles. *Biometrika*, 75, 761-771.

Charnes, A., Cooper, W. W. & Rhodes, E. (1978). Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, 2, 429-444.

Cressy, R. (1996). Are business start-ups debt-rationed?. *Economic Journal*, 106, 1253-1270.

Deprins, D., Simar, L. & Tulkens, H. (1984). Labor-Efficiency in Post Offices. In M. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The Performance of Public Enterprises: Concepts and Measurement*. North Holland: Elsevier Science Publications B. V.

Evans, D. S. & Leighton, L. (1989). Some Empirical Aspects of Entrepreneurship. *The American Economic Review*, 79, 519-535.

Foreman-Peck, J. (2005). Measuring Historical Entrepreneurship. In Y. Cassis & I. P. Minoglou (Eds.), *Entrepreneurship in Theory and History*, New York: Palgrave Macmillan.

Grilo, I. & Thurik, A. R. (2008). Determinants of entrepreneurial engagement levels in Europe and the US. *Industrial and Corporate Change*, 17, 1113-1145.

Harada, N. (2003). Who succeeds as an Entrepreneur? An Analysis of the Post-entry Performance of New Firms in Japan. *Japan and the World Economy*, 15, 211-222.

He, X. (1997). Quantile curves without crossing. *American Statistician*, 51, 186-192.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799-821.

Hurvich, C.M. & Simonoff, J.S. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, 60, 271-293.

Kieschnick, R. & McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3, 193-213.

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

Kokic, P., Chambers, R., Breckling, J. & Beare, S. (1997). A measure of production performance. *Journal of Business & Economic Statistics,* 15, 445-451.

Meeusen, W. & van den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Errors. *International Economic Review*, 18, 435-444.

Newey, W. K. & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.

OECD/Eurostat (2009). Measuring Entrepreneurship: A collection of indicators. Available at http://www.oecd.org/dataoecd/43/50/44068449.pdf.

Ray, S. C. (2004). *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*, Cambridge: Cambridge University Press.

Robb, A., Reedy, E. J., Ballou, J., Des Roches, D., Potter, F. & Zhao Z. (2010). An Overview of the Kauffman Firm Survey: Results from the 2004-2008 Data. Available at http://www.kauffman.org/uploadedFiles/kfs_2010_report.pdf.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, New York: Springer.

Wilson, P.W. (1993). Detecting outliers in deterministic nonparametric frontier models with multiple outputs. *Journal of Business & Economic Statistics*, 11, 319-323.