RESAMPLING AND BOOTSTRAP ALGORITHMS TO ASSESS THE RELEVANCE OF VARIABLES: APPLICATIONS TO CROSS-SECTION ENTREPRENEURSHIP DATA*

Jose Ignacio Gimenez-Nadal [a,b], Miguel Lafuente [a],
Jose Alberto Molina [a,b,c] and Jorge Velilla [a]

[a] *University of Zaragoza, Spain*; [b] *BIFI, University of Zaragoza, Spain*; [c] *IZA, Germany*

Abstract

In this paper, we propose an algorithmic approach based on resampling and bootstrap techniques to measure the importance of a variable, or a set of variables, in econometric models. This algorithmic approach allows us to check the real weight of a variable in a model, avoiding the biases of classical tests, and to select the more relevant variables, or models, in terms of predictability, by reducing dimensions. We apply this methodology to the Global Entrepreneurship Monitor data for the year 2014, to analyze the individual and national-level determinants of entrepreneurial activity, and compare results with a forward selection approach, also based on resampling predictability, and a standard forward stepwise selection process. We find that our proposed techniques offer more accurate results, which show that innovation and new technologies, peer effects, the socio-cultural environment, entrepreneurial education at University, R&D transfers, and the availability of government subsidies, are among the most important predictors of entrepreneurial behavior.

Keywords: Bootstrap, regression, logit, GEM data, entrepreneurship.

JEL codes: C21, C52.

* Correspondence to: J. Ignacio Gimenez-Nadal, Department of Economic Analysis, Faculty of Economics, C/ Gran Via 2, 3rd floor, 50005 – Zaragoza, Spain. Tel.: +34 876 554683. Fax.: +34 976 761996 email: ngimenez@unizar.es.

1. Introduction

Statistics is a relatively new field of knowledge that has evolved rapidly in recent decades. Advances in computing, and the recent development of new tools and frameworks such as *Bayesian statistics*, *machine learning*, and *big data*, make Statistics a very dynamic field. But despite these new tools, cross-sectional econometric analyses often make use of classical models (linear and logistic regressions) to analyze data and make inferences. While these models have the advantage of offering easily interpretable results, they rely on strong hypotheses (e.g., homoskedasticity, normality of residuals, and linear incorrelation of regressors) which makes the results and conclusions obtained from the models dependent on the fulfillment of these hypotheses. Certain authors have criticized the conclusions obtained from these methods, as they are not really drawn from the problem (i.e., from the real mechanism of the nature of the problem) but from the correct or incorrect operation of the model (Breiman, 2001b). [1] Furthermore, the usual tools when assessing the accuracy of models (e.g, BIC, AIC, R-squared, Mean Square Error) may suffer from overfitting (Friedman *et Al.*, 2001)[2]. Thus, while the latter approaches are generally accepted in economic analyses, there are compelling reasons to propose alternatives.

The goal of this paper is to propose –and apply to a known data set– an alternative approach based on the combination of well-known tools, that are resampling and bootstrapping, and to draw conclusions from cross-sectional models where some of the classical hypotheses do not have to be imposed. In doing so, we apply our approach to the Global Entrepreneurship Monitor (GEM) data, to analyze the entrepreneurial phenomenon. We conclude that our approach leads to improved models in terms of accuracy, which may make the results and conclusions drawn from them more reliable in general terms (Breiman, 2001b). Despite that we propose an algorithmic approach, developed for continuous or dichotomous variables, it can be easily extended to other frameworks, such as multiple-categorical variables, by selecting suitable alternative error functions. We contribute to the literature by proposing a simple solution to a

[1] Examples of results that depend on the fulfillment of linear and logistic regression hypotheses are the estimated coefficients, their significance, and the accuracy of the model.

[2] In this context, the term “accuracy”, or “accurate”, refers to the degree of closeness of a given model’s predicted outcomes and the corresponding true values. Nonetheless, this notion requires an error function and a corresponding estimator, as will be seen later (e.g., the mean average error of prediction).

vexing problem, that of variable selection, when there are too many variables to efficiently be used in a particular model.

In our algorithmic procedure, the predictive power of a set of variables is computed, which can then be used to estimate the proportional importance of a variable in an econometric model.[3] In doing so, we eliminate the biases incurred using classical econometric models. Our approach is based on resampling and bootstrap techniques, relies on the honest estimation of the infinite absolute test set error (also called the generalized absolute error in *machine learning*) and its comparison, and takes into account the risk of overfitting by selecting training and test sets from our samples. This method can hence be seen as an alternative to information criteria and R^2 statistics to compare models.

We then apply these methods to analyze entrepreneurial activity at both individual and country levels. Entrepreneurship is considered to be an engine of development and a source of employment, especially since the recent economic crisis of 2008, and it has been widely studied in a range of scenarios (e.g., Brown and Ulijn, 2004; Acs et *al.*, 2005; Acs, 2006; Thurik, 2009; Stuetzer et *al.*, 2012; Grimm and Paffhausen, 2015; Velilla and Ortega, 2017). Further, many Governments and institutions usually offer a range of help and support for entrepreneurs, at regional, national, and international levels (e.g., the *2020 Entrepreneurship Action Plan*). However, because of the lack of consensus about the effectiveness of these measures, and the fact that entrepreneurship is considered a very complex phenomenon, deeper analyses must be carried out in order to better understand such labor and social activity (Coduras et *al.*, 2015; Naudé, 2016).

We first apply the proposed algorithmic approach to the GEM National Experts Survey (NES) data for the year 2014 to, on the one hand, review its various applications and its ways of use, in order to extract information through data examination. On the other hand, we analyze entrepreneurial activity from a point of view that is new in this field.[4] We find that living in a European country is the most important predictor of entrepreneurial behavior. Furthermore, entrepreneurial education, the availability of

[3] The Nobel Prize-winner in Economics, Milton Friedman, suggests that the predictive power of a model is the key in determining whether relationships between variables are meaningful or not (Friedman, 1953).

[4] The GEM (Global Entrepreneurship Monitor) is “the world’s foremost study of entrepreneurship” (http://www.gemconsortium.org), and it provides to the scientific community high quality data and reports about entrepreneurial activity around the world. A recent review o GEM analyses can be read in Singer et *al.* (2015).

government subsidies, culture, Research & Development (R&D) transfers, and equity of access to new technology are the most powerful predictors of entrepreneurial behavior, as their effects are the strongest among all the variables. Furthermore, we repeat the process with the GEM Adult Population Survey (APS) Global data for the year 2014, which includes individual-level information on entrepreneurial behavior and allows us to apply our approach to dichotomous variables. We obtain that peer effects and innovation factors are the main predictors of entrepreneurial activity. In these two different scenarios, our algorithmic approach yields comparatively better results, in comparison to the classical forward stepwise selection methods.

Our contribution to the literature is twofold. First, we group some well-known statistical techniques and develop an algorithmic method to compute the predictive power of a variable within a group of variables, and then estimate the strength of its relationship to the outcome using its predictive power. Thus, we redress the issue of assessing the importance of variables through classical and biased tests (*t*-tests and their significance). The methodological novelty of our approach is not the statistical techniques used, but the fact that such methods are underused in microeconometrics, and we make use of them by aggregation, formalizing an algorithmic approach that can be used in academic settings to extract information from data. Second, we apply this algorithm to two current GEM databases, the GEM 2014 NES data and the GEM 2014 APS Global data, to analyze which variables are best related to entrepreneurial activity, and compare these results with the ones obtained from other approaches. This way, we contribute to the understanding of entrepreneurial activity, a complex phenomenon that has become increasingly important since the recent economic crisis, as a way to decrease unemployment and drive economic growth and development.

The rest of the paper is organized as follows. Section 2 contains a brief summary of the existing literature. In Section 3, we develop our algorithmic approach. Section 4 presents the data, and in Sections 5 and 6 we apply our methodology to these datasets. Finally, Section 7 presents our main conclusions.

2. Literature review

The concept of bootstrap was first proposed by Efron (1979, 1982) as a procedure called "resampling", which involves repeatedly drawing samples from a training set and refitting a model of interest for each sample. Many authors have since studied the properties of bootstrap and resampling procedures, and today it is a standard chapter in statistical learning handbooks (e.g., Friedman *et al.*, 2001; James *et al.*, 2013). Furthermore, there are several books and studies that analyze bootstrap specifically and review its various applications and perspectives (e.g., Vinod, 1993; Jeong and Maddala, 1993; Horowitz, 1997; Davidson and MacKinnon, 2006; and MacKinnon, 2002, 2006). Among the uses of the bootstrap technique, we find the following: 1) computing standard deviations of quantities of interest in difficult or complex situations; 2) quantifying the uncertainty associated with a given estimator and, related to this, 3) improving statistical learning methods (see Efron and Gong, 1983; Gong, 1986; and Freedman and Navidi, 1986, for examples of bootstrapping in regression contexts). Bootstrap techniques have been studied in a variety of settings. For instance, Breiman (1996, 2001a) applies bootstrap aggregation to reduce the bias and the variance of tree-based regressions, Büchlmann and Yu (2002) analyze it as a tool to improve the accuracy of models, and Dudoit and Fridlyand (2003) apply it to cluster analysis, It has also been applied to other fields, such as ecological problems (e.g., Quinlan, 1996; and Prasad at *al*., 2006) and genetics (Strobl *et al.*, 2007). However, to the best of our knowledge, it has been underused in economic analyses.

Within this framework, variable selection (or 'subset' selection) is an important issue in statistical problems (see George, 2000, and Guyon and Elisseeff, 2003, for an introduction to this topic) that can be improved by using bootstrap. It is usual to set three families of bootstrap methods (Kohavi and John, 1997; Guyon and Elisseeff, 2003): 'filter' (variable importance does not depend on a given model), 'wrapper' (includes the prediction performance), and 'embedded' bootstrap (combines variable selection and model estimation;). Bootstrap embedded methods may offer better results in terms of accuracy and, then, the process of variable selection would suffer less from biases and have lower variance, although they do entail high costs of computation in comparison with other techniques (Zheng and Loh, 1995).[5] Thus, the variables selected provide more accurate models. Austin and Tu (2004) bootstrap variable selection using

[5] Computational advances in recent times mean that the problem of computation costs is lessening in importance, making bootstrap techniques more prevalent.

stepwise selection based on the significance of *t*-statistics, and Veall (1992) shows an econometric example of the bootstrap when inferences are made from empirical models based on a series of trial estimates (i.e., on mined data) and classical models offer inaccurate results. However, these studies rely on *t*-statistics-based stepwise selection processes, which may also be biased because of *t*-test hypotheses. Despite that potential problem, the results derived from both studies show impressive improvements in accuracy, in comparison with classical models. In the same line, Kohavi (1995) reviews the accuracy of estimation methods based on cross-validation and bootstrap, and Rao and Tibshirani (1997) propose a bootstrap method for model selection in order to mimic Bayesian methods without a Bayesian structure and to establish a selection criterion.

Bootstrap techniques are also used in two-step econometric models where certain regressors are imputed because they are measured with sampling error, and thus second-stage tests based on a covariance matrix are biased (Pagan, 1984; Murphy and Topel, 2002). Nonetheless, little importance has been given in econometrics to empirical problems that are apparently less complicated but are common in the social sciences, such as the accuracy of regression tests on data that does not necessarily satisfy regression hypotheses, or variable selection, and that can be partially addressed with resampling techniques, since it has been theoretically demonstrated that such techniques improve the accuracy and stability of statistical (and econometric) models. Horowitz (1997) and MacKinnon (2002) each provide a review of bootstrap-based inferences in econometrics, showing how Monte Carlo tests, bootstrap tests, and confidence intervals have dominated the attention of researchers.[6] A few examples of econometric studies making a central point of the bootstrap are Horowitz (2003), who presents examples of performance of bootstrap econometric techniques; Adkins and Hill (2007) and Freedman and Peters (1984), who apply bootstrap to Monte Carlo simulations; Diebold and Chen (1996), who study structural change tests, and Li and Maddala (1997), who apply bootstrap to cointegrating nonstationary time series to significantly improve results over classical asymptotic inferences. However, econometric analyses do not usually use the bootstrap to deal with subset selection and to measure the goodness of fit of cross-sectional models, which is one of the objectives of our present study.

[6] Bootstrap tests are an alternative to Monte Carlo tests when the tests are not pivotal.

3. Description of the algorithm

In this Section, we develop an algorithmic method to measure the *importance* of explanatory variables from the point of view of their ‘predictive capabilities’. Within this framework, we propose that variables with a high predictive power are “meaningfully” related to the dependent variable, and this measure is more reliable than that of the significance of a relationship as measured through classical individual *t*-tests.

The “predictive capability” of a variable depends on the form of the problem. In the case of regression models, we make use of the *mean absolute error* (m.a.e.) of the prediction (i.e., the average of the absolute differences between real values and their estimations):

$$m.a.e. = \frac{\sum_{i=1}^{N} |Y_i \text{-} \hat{Y}_i|}{N} \tag{1}$$

where *N* is the number of individuals, Y_i the value of the outcome variable for individual ‘*i*’, and $\hat{Y}_i$ its associated predicted value. We choose this measure rather than the *mean squared error*, since its interpretation is clearer (it is measured in the same units as the dependent variable), although the selection should not affect the results qualitatively.[7] In the case of dichotomous variables, we also make use of the mean absolute error, but comparing the numeric value of the dependent variable, labeled 0 or 1, with the estimated probability of belonging to the latter class (i.e., each error will be computed with the quantity *P;* the individual does not belong to the real class).[8]

This measure of fitting, in contrast to other measures, such as the R^2, will not be calculated with the data used to estimate the model, because this may lead to overfitting. This problem appears when we estimate the goodness of fit of a model with the same set of individuals used to *train* (or fit) this model. When this problem is not taken into account, some noisy information can be taken as good information and then the

[7] Mean Square Error (m.s.e.) is generally chosen in statistics because of its differentiable properties. However, since we do not make use of these properties, we prefer to choose the mean absolute error.

[8] In the case of a dichotomous variable, we have chosen the mean absolute error rather than the success rate of classification, because for each observation, the former (a number between 0 and 1) gives us more accurate information (sensitivity and sensibility) about the performance of the model than the success-fail result (0 or 1). Also, there is, on occasion, an unbalanced distribution of the two dichotomous outcomes of interest (i.e., one of the possible outcomes has a high frequency), and thus classical econometric models predicting always the high-frequency outcome could be considered as having a very high predictive power.

goodness of fit of the model is overestimated (i.e., the R^2 strictly increases, adding independent variables to the model, although these variables have no relationship to the dependent variable). There are several statistics, widely used, that measure the goodness of fit by penalizing the number of variables. However, these settings still depend on training data (*adjusted* R^2) and even on the theoretical distribution of the model (e.g., BIC, AIC, or AICC). In the same way, prior research has shown the lack of power in goodness-of-fit tests when the alternative direction is not correctly specified (Bickel et al. 2006).

In order to avoid these problems, we divide our sample into a group used to fit ('train') the model, a *training set* (*Tr*); along with another group used to estimate the goodness of fit, a *test set* (*Te*). Then, the mean absolute error over the test set takes the form:

$$m.a.e. = \frac{\sum_{i \in Te} |Y_i \text{-} \hat{Y}_i|}{|Te|} \tag{2}$$

where $|Te|$ indicates the number of individuals in the test set. The estimated m.a.e. (over the test set) gives us information on how rightly or wrongly our model predicts *new* data. Since this method could lead to overfit of the test set, we avoid this by resampling (Friedman *et al.*, 2001; Efron and Tibshirani, 1993). See Figure 1 for a pseudo-code summarization of the algorithmic process.

We compute a bootstrap with 5,000 iterations on the following algorithm. We first randomly select a bootstrap sample (i.e., we randomly choose with replacement a subsample of the same size as the original sample), which constitutes our training set ($Tr^{(k)}$, for each iteration *k*). It is proven that, following this rule, we asymptotically obtain a training set in which approximately 63% of the individuals in the initial sample appear at least once. The remaining individuals will constitute our test set ($Te^{(k)}$).[9] We also randomly select a subset of regressors in each iteration. In our case, we specifically choose the number of introduced features in the model to be on the order of the squared root of the total independent variables, $m_2 \approx \sqrt{m}$, based on studies for similar problems by Efron and Tibshirani (1993), Amit and Geman (1997), and Ho (1998).[10] We then

[9] Given an observation, it will belong to the test set with probability $\left(1\text{-}\frac{1}{n}\right)^n$, where *n* is the number of variables. When *n* goes to ∞, this tends to $e^{-1} = 0.3678$.

[10] The chosen random selection process asymptotically guarantees, for each single variable, an equilibrium for both their own number of participations and the recombination frequency between every

train our model with the observations of the training set corresponding to the k-th iteration:

$$Y_i = \hat{f}\left(X_{1j}, \dots, X_{m_2 j}\right) + \varepsilon_i^{(k)} \tag{3}$$

for each $i \in Tr^{(k)}$, where $X_{1j}, \dots, X_{m_2 j}$ are the features randomly selected and $\varepsilon_i^{(k)}$ the unmeasured factors of the corresponding iteration[11].

For each iteration k, once the model is trained, we predict the value of the dependent variable, $\hat{Y}$, for each individual 'j' in the test set and save those values, as well as the indices of the variables included in the model:

$$\hat{Y}_j^{(k)} = \hat{f}\left(X_{1j}, \dots, X_{m_2 j}\right) \tag{4}$$

for each $j \in Te^{(k)}$. Then, the k-th iteration finishes and begins the (k+1)-th.

Finally, to estimate the overall performance of a particular variable, we average the estimated mean absolute error computed over the corresponding test set for each iteration in which it has participated. The output of our algorithm is then a vector of size the number of total variables. Each element 'i' in the vector, $\epsilon(X_i)$, contains the average of the mean absolute error (over the test set) of the models in which the i-th feature has been added, i.e.:

$$\text{output} = (\epsilon(X_1), \dots, \epsilon(X_m)),$$

where

$$\epsilon(X_i) = \frac{\sum_{X_i \in \text{Model}_k} (\text{m.a.e.}(\text{Model}_\text{k}))}{\text{interventions of } X_i} \qquad 5)$$

for $i = 1, \dots, m$, and m.a.e.(.) represent the mean absolute error over the test set of the model (equation (2)). Thus, the lower this quantity associated with a variable, the greater the predictive power of the correspondent variable, and thus the greater the importance of that variable, independently of its associated t-ratio (and the correspondent p-value). Furthermore, we should compare the differences between

subset of variables, assuring their interaction with the whole database in the presence of different regressors at each iteration. For this last point, the chosen number of regressors in the order of the square root of the total number of variables is often used as a rule of thumb, supported by empirical studies, when performing non-random-selection in statistical learning, based on the previously cited reason of interaction with a part of the data-based depending on its whole size.

[11] In equation (3) is written for simplicity Y_i as a function, although our approach can be more generally applied to non-functional relationships, as for example, when estimations are based on genetic algorithms or other kinds of stochastic modifications of the more usual statistical learning methods.

$\{\epsilon(X_k)\}_{k\in G}$ and $\{\epsilon(X_k)\}_{k\notin G}$, to have a measure of how important it is to add X_k to the model.

We acknowledge that this method allows us to compute the importance of a variable according to its predictive power (i.e., how it helps to predict the dependent variable). However, we do not obtain the sign of the relationship, nor its level of significance based on a test (e.g., Veall, 1992; Austin and Tu, 2004), which is what we aim to correct, due to the limitation of those statistics. While running this algorithm, due to the computational cost of resampling, and the time of fitting each model for large databases, we should be careful when choosing the total number of iterations, trying to find an equilibrium between this quantity and the computation-time trade-off. We also consider the possibility that the number m_2 of variables participating at each iteration is too high and then we are adding some noisy information to the model, which would lead to some confounding result (James et *al.*, 2013). The set of variables included at each iteration of the model does not necessarily provide a good model, and consequently this measure is not an estimation of the best possible performance. Nevertheless, the quantity will be useful for comparing the power of each variable when it interacts with the whole database information. In subsections 5.2 and 6.2, we propose a complementary method to deal with this issue, based on the previously-described algorithm, and compare the outcomes of both approaches with the standard forward stepwise selection process (based on t-type tests) in subsections 5.3 and 6.3.

It is important to highlight that we do not restrict this analysis to a concrete cross-sectional model (parametric, or not) or algorithm (CART, genetic algorithms…), because the measure of fit of the model, the mean absolute error over the test set, can be extended for every regression problem. An advantage of this algorithm is precisely its capability of being applied to many statistical techniques by changing the regression method or algorithm, despite that these methods could have a stochastic nature. Our methods contain many similarities with the out-of-bag error in random forest (defined originally for tree techniques, Breiman 2001a) or bagging (Breiman, 1996) methods. Nevertheless, out-of-bag estimations are defined for a single model in terms of the bagged model using different loss functions (m.s.e. and rate of failure for regression and classification problems respectively). In addition, since we focus on inference rather than prediction, we choose a more suitable error function, especially in the dichotomous variable setting where the 0-1 loss may be unable to capture some information in problems where the relative frequency of groups is unbalanced.

Thus, the algorithm allows us to reliably compare different kinds of model, not necessarily of the same type and specification, and choose the best one in terms of its predictability power, but with the aim of making the most reliable inferences, which is the objective of economic studies. For example, if we use logistic regressions, we do not need to assume that every hypothesis is true. However, we focus on the degree of prediction of the model and draw conclusions from it, rather than from *p*-values or *t*-ratios that are probably biased because the data may not follow the required hypotheses.

4. Data and variables

We first use the Global Entrepreneurship Monitor (GEM) National Level data for the year 2014, in order to analyze which socio-economic characteristics of countries are related to entrepreneurial activity, which is measured by the TEA index.[12] This database is annually elaborated by GEM and contains information about nine groups of variables of national, socio-economic characteristics related to entrepreneurship encouragement (a description of those variables can be found in Table 1). This data is based on surveys filled out by 36 experts in each country, where stylized questions are used. For every country, each of the 56 variables takes the mean average responses of the experts, with values between 1 (totally disagree) and 5 (totally agree). More information can be found at http://www.gemconsortium.org/about/wiki. Since we have data about every single characteristic, and also about each group, or principal factor, of variables (see Table 1), we will repeat the analyses for both cases.[13] We have information on 69 countries: Australia, Austria, Belgium, Canada, Chile, Denmark, Estonia, Finland, France, Germany, Greece, Salvador, Slovakia, Slovenia, Spain, Hungary, Ireland, Italy, Japan, Luxembourg, Mexico, Norway, Netherlands, Poland, Sweden, Switzerland, United Kingdom, United States, Angola, Argentina, Belize, Bolivia, Bosnia, Botswana, Brazil, Burkina-Faso, Cameroon, China, Colombia, Costa Rica, Croatia, Ecuador, Filipinas,

[12] The TEA (Total early-stage Entrepreneurial Activity) index measures the percentage of the working-age population who have begun a new business in the recent two and a half years, or intend to do so. This indicator is computed yearly by GEM and is downloadable from http://www.gemconsortium.org/data/sets.

[13] The GEM NES National Level already defines every factor, as shown in Table 1. For a given country, each principal factor value is the mean of the single variable values associated with it.

Georgia, Guatemala, India, Indonesia, Iran, Jamaica, Kazakhstan, Kosovo, Lithuania, Malaysia, Panama, Peru, Puerto Rico, Qatar, Romania, Russia, Singapore, South Africa, Suriname, Thailand, Trinidad and Tobago, Uganda, Uruguay, and Vietnam.

Summary statistics of variables and principal factors are seen in Table 1, showing that the average TEA index across countries is 13.08, which means that in the mean country, 13% of the working-age population would be entrepreneurs. Further, the standard deviation is 8.15, indicating a considerable variation across countries. Furthermore, 33.3% of the countries in the sample belong to the EU, and 42% to the OECD. The rest of the variables are, in general, between 2 and 3 (in the scale of agreement), with standard deviations rarely greater than 0.5. The lowest values are found among variables of the group "Entrepreneurial level of education at Primary and Secondary", with a mean value of the factor of 2.09 and a moderate standard deviation of 0.39. On the other hand, the highest values are reached in the groups "Professional and commercial infrastructures access" (3.02, with a moderate standard deviation of 0.33) and "Physical infrastructures and services access" (3.75, with a standard deviation of 0.48, higher than in the previous case). A recent analysis of the entrepreneurial level of countries using GEM data can be found in Molina and Barrado (2015).

We also use the GEM APS Global Individual Level database of the year 2014 to analyze the individual factors related to the fact of being an entrepreneur, or not. The database is elaborated by GEM on an annual basis and contains information about entrepreneurial-related characteristics, and is defined at the individual level (microeconomic data). This data is based on stylized questions elaborated by GEM experts, and filled-in by individuals in each of the analyzed countries. Each question is answered between 1 (totally disagree) and 5 (totally agree). Thus, we are able to analyze entrepreneurship from the point of view of individuals, and macroeconomic factors.

We now have data on 188,373 individuals, 93,636 of whom belong to an OECD country and 94,738 of whom do not. Summary statistics of the 53 features and the dependent variable are shown in Table 2. We also have information on the country of residence of each individual, information that will be taken into account in the analysis. In order to make the analysis less susceptible to biases, we have redefined variables as *dummy*, taking value 1 if the answer is an agreement (values 5 as "totally agree" or 4 as "agree"), and value 0 for the remaining categories.

We can see that only 8.1% of individuals in the OECD are entrepreneurs, in contrast with 16% in the non-OECD countries. The difference is significant, according to the Kruskal-Wallis test. In fact, all characteristics taken into account are significantly different between individuals living in OECD and non-OECD countries, except for the perception of the conditions of life. Some of the most meaningful differences can be found in: 1) age (with an average of 43 years in the OECD and 38 in the non-OECD countries); 2) the consideration of the own skills of the entrepreneur (55% of the individuals in non-OECD countries think they have sufficient skills to be an entrepreneur, in contrast with 44% of OECD individuals); 3) the proportion of individuals who are paid-employed is 54% in the OECD, vs 39% in non-OECD countries, which can have important effects on those who are entrepreneurs due to necessity; 4) the desire to become a businessman is present in 66% of the individuals in non-OECD countries, in contrast with 44% in the OECD who wish to become an entrepreneur; 5) the consideration of finding good opportunities to become an entrepreneur (43% vs 26%); 6) the good social perception of being an entrepreneur (44% vs 26%); 7) the consideration of an entrepreneurial supporting culture (66% vs 42%); and 8) the promotion of entrepreneurs through the media (64% vs 41%).

5. An application to continuous variables

5.1 First step

Following the algorithmic approach in Section 2, we consider the TEA variable for the GEM 2014 NES National database, by country, as our dependent variable, and the rest of the variables as regressors. In each iteration k, we consider the following linear regression model:

$$Y_i = \beta_0^{(k)} + \sum_{i=1}^{m_2} \beta_i^{(k)} X_i + \varepsilon_i^{(k)}, i \in Tr^{(k)} \qquad (6)$$

with $m_2 = 7$, the number of variables in each iteration, and ε_i being unmeasured errors with a normal centered i.i.d. distribution, obtaining the results shown in Figure 2 (with a computation cost of 8.904 seconds for 5000 bootstrap iterations).[14] In other words,

[14] m_2 is the number of variables in each iteration, and we choose it to be around sqrt (total number of variables), as mentioned in the previous section. As we have 56 variables, we take m_2 =7. We use the open source and free software R, with an intel i7 CPU running at 4.0GHz, to develop the applied part of the study.

Figure 2 represents, for each exogenous variable (represented on the X axis) the average value of the mean absolute error (over test sets) associated with the dependent variable (Y axis), considering all iterations in which the corresponding variable appears, which is estimated following the algorithmic approach. That is to say, for every feature on the X axis, we estimate equation (5) and then represent it on the Y axis. Thus, we must interpret that the smaller the value of the m.a.e. associated with a given feature, the better is the model after its inclusion (i.e., the less important the error committed, and the better the prediction).

In Panel A, the process is developed, taking into account all single variables, excluding the principal factors. We find that belonging to the European Union is clearly the most powerful regressor, perhaps indicating the structural characteristics of this group of countries that condition their entrepreneurial level. However, the variable "country" appears to be the less influential regressor, indicating that including country fixed effects, apart from the indicators taken into account in the model, may not contribute in a meaningful way to the accuracy of the models.

Variables D2.4 ("Colleges and Universities provide good and adequate preparation for starting up and growing new firms"), C.1 ("A wide range of government assistance for new and growing firms can be obtained through contact with a single agency") and A.3 ("There are sufficient government subsidies available for new and growing firms") also contribute clearly to model accuracy, in contrast with the rest of the regressors. Thus, our results indicate that entrepreneurial education at University, simplicity of Government assistance, and availability of Government subsidies are among the factors with the greatest impact on the entrepreneurial activity of countries. Variables I.2 (entrepreneurial culture), G1.2 (intra-industry market dynamism), A.4 (funding availability from private investors), G2.5 (restrictions on new firms from established firms), D2.6 (vocational and professional entrepreneurial education) and I.1 (success-through-effort culture) are, apart from the latter, the independent variables with the highest predictive power, according to our methodology.[15] On the other hand, variables D1.3 (entrepreneurial education at primary and secondary levels), G1.1 (goods and services market dynamism), B2.7 (cope with bureaucracy regulation and licensing) and E.2 (equity in the access to technology) appear to be the variables with the lowest predictive power.

[15] GEM's definition of each variable is shown in Table 1.

When we focus on Panel B (where the analysis is done using the grouped factors, rather than all the single independent variables, and setting $m_2 = 3$), we can see that, again, belonging to the European Union is by far the most important variable, despite that country fixed effects are among the less important variables, which is consistent with our previous findings. The rest of the results are also in line with those found in Panel A, since entrepreneurial education at University (group D.2 in Table 1) and socio-cultural support (Group I) are among the five variables with the higher predictive power, together with R&D Transfers (Group E) and belonging to the OECD. In contrast, the three variables with the lowest explanatory power regarding the entrepreneurial level of a country are bureaucracy and taxes (Group B.2), country fixed-effects, and Government policies and support (Group B.1).

Our results indicate that entrepreneurship appears to be mainly motivated from vocation, desire, and social norms (Kotsova, 1997; Minniti, 2005; Cooper and Yin, 2005; Arenius and Minniti, 2005; Terjesen and Szerb, 2008; Molina *et al.*, 2015), where government investment and subsidies can be an incentive for individuals to initiate a business and become an entrepreneur (Acs, 1992; Lundstrom and Stevenson, 2002; Amorós *et al*, 2012; Berrios-Lugo and Espina, 2014) but bureaucracy and tax restrictions do not discourage them (in contrast with Kotsova, 1997). Furthermore, specific entrepreneurial skills education at University and the professional level are also important determinants for future entrepreneurs (Kotsova, 1997; Minniti and Nardone, 2007; Bosma *et al.*, 2004; Levie and Autio, 2013), together with concrete issues at early stages of education, where it is important to promote creativity and self-sufficiency (in line with the notion of Kyrö, 2015, that entrepreneurial education is a new form of pedagogy).

5.2 Step by step selection

In this subsection, we modify the previously-described algorithm in order to obtain complementary information. For this purpose, we look to find a subset of variables leading to the best possible prediction accuracy via a step-by-step approach. For the first step, we analyze every explanatory variable one by one, with 2,000 iterations per variable (at a computation cost of around 55 seconds). Our objective is to check, in a simple linear regression model, which variable has the greatest predictive power over test sets via bootstrap samples, as above. After that, we repeat the procedure, at each step setting the previously-selected variables and a loop to repeat the first step process (without the already-selected variables). The process ends when the average m.a.e. of the best variable included is worse than the previous one, indicating that its inclusion does not suppose any improvement in the model. In order to have reliable results, we choose a large number of iterations regarding the size of the database, due to the stochastic nature of our algorithm. Because of the computational cost of fitting models in large dimensions, the step selection should be performed forwards. This problem arises, for example, in the next section, for individual data, where a backward approach would be practically useless.

Figure 3 shows the results, where the X-axis shows the variables, and the Y-axis the average of the m.a.e. associated with each variable. We can see in Figure 3, Panel A, that the dummy EU is the first variable to enter into the model. It is important to note that the mean average error in the model whose only regressor is belonging to the EU is 5.35, almost equal to the mean average error of the 'best' variable of the previous subsection, which leads us to consider that most of the information in the database is noisy with respect to the TEA index. The second step results are shown in Figure 3, Panel B. We see that the most predictive model is the one with the EU dummy and the R&D level of transference as regressors of the TEA index, with an error of less than 5.1, which supposes an increment of predictive power with respect to the previous case. Including a third variable, we find that the best feature would be the entrepreneurial level of education at vocational, professional, College, and University (Figure 3, Panel C), giving us a new estimation of the mean absolute error of prediction of 5.00. At this point, adding more variables could result in adding noisy information, as can be seen in Panel D of Figure 3, where the best variable to add would be the financial environment related to entrepreneurship, but its inclusion supposes an increase in the mean prediction error in comparison to the previous model.

Thus, we find that the final model for the TEA index, taking into account overfitting, and considering the principal factors from the GEM database as regressors, is the one including the dummy variable for belonging to the EU, the R&D transfers (Group E), and the level of vocational, professional, College, and University entrepreneurial education (Group D.2). This result should be taken into account, since adding more variables could result in adding noisy information. Since econometric models have the goal of providing the most adequate conclusions, in terms of the economic theory, predictability should play a primary role because it gives researchers an idea of how accurate their conclusions are, relative to the data.

Figure 4 shows the results of the analysis when all variables are included, rather than only the principal components. We find that the variables that should be included in the model are, in order of appearance in the algorithm: belong to the EU (Panel A), A.3 (government subsidies for new and growing firms, Panel B), D2.4 (entrepreneurial preparation in Colleges and universities, Panel C), D1.1 (creativity, self-sufficiency and initiative in primary and secondary education, Panel D) and E.2 (new and established firms have as much access to new research and technology, Panel E); including I.3 (national culture encourages entrepreneurial risk-taking, Panel F) should be the following choice, but it increases the mean prediction error, and we conclude that the information contained in this variable is of little use in the presence of the previously-mentioned variables.

5.3 Comparison with a standard forward stepwise selection process

In order to see how the latter results may differ from the standard forward stepwise selection process (FSS), which builds models by adding the "most statistically significant" explanatory variables (based on *t*-tests), in this subsection we develop this process and compare it with the two models developed in subsections 5.1 and 5.2.

The model derived from the FSS, when the explanatory variables are the principal components, is formed by the following features, by order of selection for the process: 1) belong to the EU; 2) R&D transfers; 3) Entrepreneurial education at Universities; and 4) Commercial infrastructures. For the process developed in 5.1, we choose the first 5 variables, belong to the EU, entrepreneurial education at University, socio-cultural support, R&D Transfers, and belong to the OECD. In this case, possibly helped by the fact of having a small database, the three models are very similar. Indeed, we can observe how belong to the EU, entrepreneurial formation at universities, and R&D transfers appear in the three models obtained following the three approaches, while the remaining variables differ, leading to different performances. Moreover, we can see how the model obtained in Section 5.2 is contained in the model of both the FSS and Section 5.1, leading to nested models. The only difference between the model of 5.2 and the FSS model is one extra variable which, even when its *p*-value in the Student Individual Nullity test is lower than 0.05, our procedure found does not improve its accuracy in its presence.

When we examine those performances using the classical measures (AIC, BIC, R^2 and adjusted-R^2), and the proposed mean average error over test sets, we can check that the former are prone to overestimate the performance of models. Table 3 shows these values for the FSS model, and the two models obtained in the previous subsections.

We can see that, in terms of the four classical statistics, the FSS model would be the best of the three models analyzed. However, those magnitudes may suffer from a potential source of biases, as explained above. Oppositely, when we measure the goodness of the models by its prediction power over test sets, we find different results: the best model would be the one derived from subsection 5.2 (which has been designed to maximize this quantity). The second model in terms of predictability is the model of subsection 5.1, followed by the FSS model. Remembering that the error is measured in percentage points, the three estimations are not very far from each other, due to the similar specification of the model. At this point, it is important to note that, while the FSS model contains the model of subsection 5.2, its accuracy is lower, even when the only change has been to add one variable. This confirms, first, the biases that emerge from the classical statistics of goodness of fit, since all of them point to the forward stepwise model as the best one. Second, these results reveal that the variables chosen by their predictive power may differ from the ones selected by its significance, with the latter approach being more prone to overfit models and add noisy information, as shown from the good performances according to the classical statistics over training sets, but poorer accuracies over test sets.

Repeating the analogous process over the model of the TEA as a linear function of all the individual features, we find analogous results. According to the FSS, the model includes: 1) belong to the EU; 2) accessibility to technologies (E.2); 3) entrepreneurial preparation at Universities (D2.4); and 4) taxes are not a burden (B.5). In contrast with the previous case, the number of features is lower than in the correspondent model obtained in 5.2, although they are not nested models. For the model of 5.1, we choose the following variables: belong to the EU, Entrepreneurial preparation at Universities, simplicity in government assistance, and availability of government subsidies.

Table 3 shows the AIC, BIC, R^2, adjusted-R^2, and mean absolute prediction errors over test sets of this model, and of the correspondent models obtained in 5.1 and 5.2. Once more, the classical measures point toward the FSS model as the best one. However, in terms of the mean average errors over test sets, the most accurate model is again the one developed in 5.2, followed by the model in 5.1, and the FSS model.

Additionally, it is important to note the differences found between the procedure presented in Section 3, and applied in Subsection 5.1, with the step-by-step approach developed in subsection 5.2. In particular, the former shows a mean absolute error over test sets of 5.223 in the case of the principal components (and 5.013 in the case of the disaggregated features), vs 5.030 (4.982) of the step-by-step model. Since the latter approach is designed to minimize the noisy information arising from the explanatory variables of the model, and the former is built with the best variables when interacting with the whole database, such differences can give us a measure of the amount of noise between both procedures. In this particular case, in which the outcome is measured in percentages, we could conclude that the amount of noise that emerges from the algorithmic models developed in Section 3 is negligible in the case of the individual variables' analysis, and higher but still not meaningful in the case of the principal components' analysis. Nevertheless, they are not intended to be compared, but to be read at the same time in order to make inferences. Also, we can see how these issues related to noisy information are not captured by the classical measures employed, which indicate that the step-by-step model would be the worst model.

6. An application to dichotomous variables

6.1. Extension of the main algorithmic approach

Although in this section the dependent variable is categorical (dichotomous), the same main ideas apply. As previously described, our methods not only allow us to find the variables with the strongest relationships in terms of their predictive power, but also our chosen measure of goodness of fit allows us to compare different kinds of model. We will compare the performance of a logistic regression model and a linear discriminant analysis (LDA) model in predicting the probability of the contribution to the TEA index outcome, making use of the GEM 2014 APS Global database.

We predict the probability of the dependent variable taking the value 1 (i.e., the individual is an entrepreneur, 0 otherwise) and, for each feature and each iteration k in which it participates, analogously as in subsection 5.1., we keep the mean of the absolute value of the difference between the real values of Y and the predicted probability of being an entrepreneur, $\hat{Y}$, as follows:

$$m.a.e.(\mathrm{model}_k) = \mathrm{mean}\left(\left|Y_j^k\text{-}\hat{Y}_j^{(k)}\right|\right) \tag{7}$$

for each $j \in Te^{(k)}$.

Then, we take for each variable the mean of the $m.a.e.(\mathrm{model_k})$ for all the iterations in which it participates (Equation 5), obtaining an honest estimation of the mean absolute prediction error of the correspondent variable.

In the logistic regression model, for an individual 'i', we assume that the dependent variable Y_i follows a Bernoulli distribution with parameter p_i, where we must estimate the coefficients of the following equation:

$$logit(Y_i) = \ln\frac{p_i}{1-p_i} = \beta_0 + \sum_{k=1}^{K}\beta_k X_{ik} + \varepsilon_i \tag{8}$$

where $\{X_k\}_{k=1}^{K}$ are the regressors and ε_i is the error term. This expression gives us a final estimation of the probability of belonging to the class labeled as 1, as follows:

$$p_i = \frac{e^{\beta_0+\sum_{k=1}^{K}\beta_k X_{ik}}}{1+e^{\beta_0+\sum_{k=1}^{K}\beta_k X_{ik}}} \tag{9}$$

On the other hand, LDA, sometimes called the Fisher Discriminant Analysis, assumes the distribution function of each target group to be multivariate normal. Linear discriminant analysis also assumes the covariance matrices to be equal in every group (homoskedasticity), in contrast with Quadratic Discriminant Analysis. In this way, the density of each group k (in our case k={0,1}), is supposed to be *bell-shaped*

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(\text{-}\frac{1}{2}(x\text{-}\mu_k)^T\Sigma^{-1}(x\text{-}\mu_k)\right), \tag{10}$$

where p is the number of regressors, ^{T} is the transposition operator, μ_k is the vector of means of dimension p for the k-th group, and $\sum$ the covariance matrix, which is equal across every group, as pointed out above. Under these assumptions, Bayes' Theorem gives us the estimated probabilities of belonging to a group k for a particular individual with attributes x as follows

$$p_k(x) = \frac{\pi_k * \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x\text{-}\mu_k)^T\Sigma^{-1}(x\text{-}\mu_k)\right)}{\sum_{l=0}^{1} \pi_l * \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x\text{-}\mu_l)^T\Sigma^{-1}(x\text{-}\mu_l)\right)}, \tag{11}$$

with π_k being the prior probability of each group. Thus, in order to fit this model, we need to know the *a priori* probabilities, the mean vectors, and the covariance matrix for each group. As is usual in this setting, mean vectors and the covariance matrix can be then estimated by maximum likelihood, while prior probabilities can be estimated with the relative frequency of each group in the training set.

Finally, we remark that this technique is called *linear* because the decision rule to classify an individual in any particular class has a linear form. Performing some algebra, we arrive at the following discriminant function or decision rule:

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k\text{-}\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln(\pi_k), \tag{12}$$

and consequently, we will assign the observation with features x to the class in which $\square_{\mathrm{k}}(x)$ is larger. As a consequence of this result, LDA analysis assigns a single coefficient to every variable, with these coefficients being the final weight in the decision rule, and describing the effect of the regressors on the response.

Figure 5, Panel A shows the result of applying our algorithm to a logistic regression model, making use of the GEM 2014 Individual data with 500 iterations (with a computation cost of 1286.176 seconds). The interpretation of this figure is analogous to the interpretation of Figure 2. For each variable (on the X-axis), we plot the average of the m.a.e. (Y-axis), considering all the iterations in which the characteristic appears. Then, the smaller the mean error, the better the prediction and the more important is the corresponding feature. We find that the most important variables (and then the variables with the greatest effect according to their predictive power) are searching for a new market (index 23), having recently helped other entrepreneurs (index 9), can offer a new product and new technologies (indexes 10 and 11, respectively), and being a businessman or self-employed (indices 118 and 14). In particular, we find strong and unbiased empirical evidence of the importance of oral and meeting transmission as a determinant factor of entrepreneurial activity (in line with Holcomb *et al's.*, 2009, entrepreneurial learning by seeing, and Blumberg and Pfann's, 2015, parent's transmission). We also find evidence of the importance of innovation as a determinant of becoming an entrepreneur (Schumpeter, 1934; Gilbert, McDougall and Audrestch, 2006).

With the same number of iterations, we repeat the analysis, applying the algorithm to an LDA model. Results are shown in Panel B of Figure 5. Although the three former variables in the previous paragraph appear again among the most important variables, we find that the order is different, in that “have helped others to entrepreneur” now appears to have a stronger effect than “being self-employed”. Furthermore, “being a businessman” is not among the five most important variables (it is the sixth), in favor of the variable “new market”. The economic explanation behind these results is essentially the same as in the previous case.

In view of our results, we note that variables that tend to decrease the prediction error are the same, in general, for both models. The fact of them being important variables, regardless of the initial hypothesis, gives us strong evidence of the correctness of our conclusions.

Despite the remarked similarities, a significant difference between the logistic regression and the LDA models applied to our data is the scale of the error estimations. The minimum values for the former are around 0.09, while minimum values for the latter are around 0.06. This is a large difference, given that it is a difference in probabilities, and the baseline error (i.e., the error of the simplest model that predicts for all individuals the most common value of the outcome variable) is 0.122. Then, given these results, it seems that, for this particular case, the LDA model outperforms the logit model. Nevertheless, the most interesting point is the strong empirical evidence provided by the similarities in the chosen variables independently of the model or the hypotheses.

6.2. Step-by-step selection and some remarks

Econometric models in the field of micro-econometrics for dichotomous variables tend to include many explanatory variables, and the problem of overfitted models may consequently have a strong presence in these analyses. In order to select the most important variables and avoid this problem, we apply again a step-by-step forward selection from the GEM 2014 Individual data. As noted above, due to the size of the database, a backward selection would take too much computing time. Thus, we analyze the prediction error for the first steps of a forward stepwise selection with 10 iterations for each step and each variable (with computation costs of around 500 seconds per step).

Figure 6, Panel A shows results for the logistic regression model, and Panel B for the LDA model. We represent on the Y-axis the corresponding estimation of the m.a.e. over the test set for the best sequence found by the step-by-step procedure, with the indices of the X-axis being the total amount of variables at each step. Then, the smaller the associated average m.a.e. associated with a given variable, the better the model, including this variable and all previous ones.

We find that, for the LDA case, the ten 'best' variables are, in order: "have helped other entrepreneurs", "have new technologies available", "belong to a new market sector", "being a businessman", "age", "being an active unemployed", "being satisfied with the current job", "living in Israel", "consider to have opportunities to entrepreneur", and "living in China". Among our estimations, we get the minimum generalized absolute error of predicted probabilities of around 0.046 with the six first variables. Adding more variables does not improve the model in terms of the prediction error, indicating that these six offer us the best performance. Furthermore, independently of the possible existence of a *statistically significant* relationship between the rest of the potential regressors and the output variable, its inclusion in the model does not add useful information. Multiple factors can lead to this phenomenon, with this behavior for the *p*-values being very common, especially in large databases. For example, the existence of multicollinearity with the six included regressors, possibly combined with a fit of noise when dealing with the whole database. This last factor is often related to an inadequate hypothesis in practice.

We can also discern a significant difference in the error term when we include the fourth variable, but only minor differences with the addition of the fifth and sixth. Thus, its inclusion may slightly complement the information offered by the four previously-added variables. In any case, we remark that after the sixth variable the estimation of the mean absolute error over the test set tends to be flat for an increasing number of variables. In fact, this is a typical phenomenon in statistical learning methods when assessing accuracy. We can observe a characteristic U-shape for the estimated test error, which, in this case, due to the large number of individuals in the database, is attenuated and appears flat. This behavior contrasts with the estimated error in the trained set, which is always decreasing when the flexibility of the model, in this particular case the number of variables, increases. The relation between the behaviors of the two kinds of error is part of a wider issue present in every problem of supervised learning, called *the bias-variance tradeoff*.

These results indicate that, apart from the fact that the variables mentioned in the previous paragraphs are important determinants of entrepreneurial activity, age should always be taken into account in entrepreneurial empirical models (Davidsson, 1989; Blanchflower, 2000; Schott and Bagger, 2004; Kelley, 2009). Furthermore, the fact of being an active unemployed should also be considered. This is directly related to entrepreneurship due to necessity: individuals who are unemployed and have no income may decide to start their own business as an alternative.

If we now focus on the logit model results (Panel A, Figure 6), we find that the ten variables that the forward stepwise algorithm has included in the model are, in order: "have helped other entrepreneurs", "belong to a new market sector", "age", "being a businessman", "having ended a business recently", "have new technologies available", and "living in Botswana", "Colombia", "Malawi", and "Vietnam" (all of these are non-developed countries). Although the list of variables included has meaningfully changed, those in the first part, i.e., the ones that should really be taken into account before adding useless information to the model, have changed hardly at all. Note that now the optimum number of variables to add should be five (since the inclusion of the sixth increases the error estimation), whereas the first four features are among the five best in the LDA analysis. It should also be taken into account that the scale of the error term is again larger in the case of the logit model than in the LDA model. Some details about the final chosen logit model can be found in Table 5.

Again, for the logistic regression model, the estimated mean absolute error becomes stable after the inclusion of the first five variables. Furthermore, Panel A shows that the following variables to be added to the model are the dummy variables that define the country of residence of individuals. This confirms the results of Figure 5: once these groups of variables are taken into account, the addition of more variables may be just adding noisy information to the model, independently of it being a logistic regression or a LDA, supporting the correctness of our conclusions. Then, once the variables "have helped others to become entrepreneurs", "belong to a new market sector", "age", "being a businessman", "having ended a business recently", "being an active unemployed" and "have new technologies available" are included in the model, the inclusion of other variables overfits the model and conclusions drawn from it may then be biased. Thus, we conclude that, although a bi-dimensional analysis could show strong relationships between other variables and the contribution to the TEA, its contribution to the prediction weakens in the presence of the cited variables.

As final remarks, we see the effect of the variable “can offer a new product”. In Figure 5, we have found this to be one of the most important regressors. However, it does not appear as a chosen variable for the model as a result of the forward selection (in Figure 6). This is a perfect example to illustrate why it is important to take this second phase of stepwise selection of variables. According to Figure 5, this innovation variable is one of the most notable regressors. However, from Figure 6, we can conclude that the information about the TEA index that this variable provides is previously contained in other variables, which are already in the model (because the relationship of the combined variables chosen with the TEA is higher). This example is intuitive, because there is a significant difference between the error terms of the ‘best’ variables and the rest. However, in other databases, where this difference is not as pronounced as here, it may happen that a variable with a high predictive global power, but which is correlated with some others of greater importance, cannot be easily found. An alternative could be to run our algorithm again with some variables fixed (those that we have decided to include following the forward stepwise procedure, prior to a high ‘decreasing jump’ in the error terms), and others randomly selected. Then, we could check whether this ‘decreasing jump’ in the error term could also be attributed to variables other than the ones already included in the model.

6.3. Comparison with a standard forward stepwise selection process

Analogously to Subsection 5.3., we now compare the logistic regression models of the contribution to the TEA obtained in 6.1 and 6.2 with the model that derives from the standard forward stepwise selection process. In particular, the latter results in a model in terms of the following independent variables: 1) know other entrepreneurs, 2) opportunities to entrepreneur, 3) high entrepreneurial perception, 4) being a businessman, 5) entrepreneurial skills, 6) have helped other entrepreneurs, 7) can offer a new product, 8) can work with new technologies, 9) being self-employed, 10) age, 11) reside in a not developed country, 12) aim to entrepreneur in the future, 13) have invested in other entrepreneurial projects, 14) being male, 15) being satisfied with income, 16) family size, 17) fear of failure, 18) live in an OECD country, 19) being unemployed, 20) have good conditions of life, 21) have freedom at work, 22) being a student, 23) have obtained important things in life, 24) have a non-stressing job, 25) being an employee, 26) like current job, 27) low income level, 28) being satisfied with current job, and 29) have stopped a business recently, plus several country dummies, making a total of 88 variables.

For the logistic model of 6.1. we choose the first 6 variables, searching for a new market, having recently helped other entrepreneurs, can offer a new product and new technologies, and being a businessman or self-employed, since after them we observe a notable jump in the estimated m.a.e., as can be seen in Figure 2. Finally, we recall that the model derived from 6.2 make use of 5 regressors and its estimation can be consulted in Table 5.

Table 4 shows the AIC, the BIC. and the pseudo-R^2 for these models, as well as the estimated mean absolute errors over test sets. The AIC, BIC, and pseudo-R2 indicate that the FSS model would be better than the models derived from the resampling approaches that have been developed. Usually, the AIC and BIC are used to sort models by goodness of fit, and it tends to be difficult to gain a precise idea of the quality and proportional performance between models by these two quantities. Pseudo-R^{2,} on the contrary, offers a more intuitive interpretation when comparing different models.

Nevertheless, the performance of these models in terms of accuracy shows a completely different scenario. The estimated generalized error decreases by more than 0.02 for both models in subsections 6.1 and 6.2 (0.061 and 0.057) with respect to the FSS model (0.081). 0.02 actually supposes a large improvement, considering the baseline error (the probability of being an entrepreneur, according to our sample, 0.122), the FSS estimation, and the units (probability and misclassification). Moreover, this difference is achieved with a very pronounced difference in the number of regressors between models, which also makes our models preferable from the point of view of simplicity (or Occam's Razor).

One variable with a clarifying behavior for the suitability of our techniques is that of "belong to a new market". In Table 5, we can see that the associated *p*-value for this mentioned variable is around 0.75, which in practice is usually viewed as a very high value, perhaps indicating a weak effect on the model. In addition, in a simple logit model with the "belong to a new market" variable being the only feature among the regressors, its associated p-value rises to 0.77 and the estimated coefficient yields a value of 21.93. We also know that the best possible performance for a logistic regression model without any regressor can be achieved by assigning the majority class for every individual, giving us a baseline rate of failure in classification of 12.2%. However, including in the analysis the new-market variable, which is supposed not to be a good predictor according to its *p*-value, via resampling we obtain an estimated rate of failure in classification of 8.447%, which supposes a significant improvement in the reduction of the considered error. This kind of error decreases to 4.239% when every selected variable in the stepwise selection is added to the model.

Comparisons are also interesting when considering the LDA model. In subsection 6.2, we found an LDA model with only six variables, which achieves an estimated m.a.e. of 0.046. This represents a substantial improvement over our logit models, and almost halves the error in the case of the FSS logit model, with a reduced number of variables. From the mathematical point of view, apart from the problem of overfitting risk (Pseudo-R^2), these large differences are due to the misspecification of the logit model. AIC and BIC depend on the number of variables and the likelihood of the model. But the standard stepwise selection is also based on the likelihood, and this likelihood, on the model hypotheses. So it is not surprising that these results tend to agree with each other. On the other hand, large databases with real (non-simulated) data are usually far from mathematical hypotheses. In our particular case, the data fits better in the LDA specification. This causes both the difference in performances and the bad behavior of the likelihood-dependent measures.

Even when the LDA model outperforms the logistic regression in this setting, we acknowledge that the variables that are really important are the same, disregarding the initial hypothesis. This is the reason why the best models found for the logit and the LDA use essentially the same variables, and are, indeed, the variables in which we are most interested in order to extract useful information about the entrepreneurship problem.

7. Conclusions

This paper proposes a way to analyze the predictive power of regressors in cross-sectional data models. We make use of well-known resampling and bootstrap techniques to estimate the importance of variables from their predictive power, based on the mean absolute error over test sets. Against this background, the risk of overfitting is mitigated and the inherent biases of the classical measures and significance tests are also avoided. Thus, the implemented algorithm allows us to alternatively and more accurately analyze the predictive power of a model (or a set of models) and the real weight of the variables in those models. We then apply this procedure to two GEM 2014 databases (NES National and Global Individual Level) to analyze which macro- and micro-economic variables are more helpful in predicting entrepreneurial activity. We apply the algorithm to linear regression, logistic regression, and linear discriminant analysis models.

Our work contributes to the literature, first, by offering an algorithmic approach, based on known statistical tools, to check the importance of variables, and also to reduce dimensionality by identifying the best and worst variables in terms of accuracy. Furthermore, we compare results with a one-by-one selection of variables, following the same algorithmic procedure, in order to detect multicollinearity and avoid overfitted models, minimizing the amount of noisy information. One limitation of our contribution is that these techniques are developed to be applied to cross-sectional models, and the application to panel data analyses would constitute a different setting, that we leave for future research.

The second main contribution arises from the application of the developed tools to GEM databases to show, first, the strength of innovation and research, and of entrepreneurial education (at the professional, vocational, College, and University stages) as factors determining entrepreneurship. Second, taxes and bureaucracy appear not to be a burden for entrepreneurs, which agrees with the hypothesis of Molina et *al.* (2015) that entrepreneurship is a vocational activity. Subsidies and Government programs supporting entrepreneurship appear not to be effective measures in overall terms, although certain of their characteristics, such as ease of access, may have a strong influence. Third, at the individual level, innovation factors, and having contact with other entrepreneurs in the past (e.g., knowing other entrepreneurs, helping other entrepreneurs in their work activities) are among the most important determinants of participation in the TEA index.

Thus, our results should be taken into consideration in order to develop future research about entrepreneurship. Furthermore, European politicians are currently using entrepreneurship as a way to improve economic growth and reverse the negative consequences of the recent economic crisis. In doing so, they are investing large amounts of money in promoting entrepreneurial activity. However, as argued in Naudé (2016), it is not clear whether the effect of entrepreneurship on economic growth is significant, or not. There is evidence suggesting that those policies that set out to provide incentives to entrepreneurship are not particularly effective and efficient (Nagler and Naudé, 2014; Karlan and Valdivia, 2011). Because of this, policy makers should take into account these present results as a guide to a more efficient promotion of entrepreneurship.

COMPLIANCE WITH ETHICAL STANDARDS

This paper has benefited from funding from the Spanish Ministry of Economics (Project ECO2012-34828). Specifically, ML acknowledges the support from project MTM-2014-53340-P. ML is member of the research group “Modelos Estocásticos”, supported by DGA and the European Social Fund. The authors declare that they have no conflict of interest.

REFERENCES

Acs Z (1992) Small business economics: A global perspective. *Challenge* 35: 38-44.

Acs Z (2006) How is entrepreneurship good for economic growth? *Innovations* 1: 97-107.

Acs ZJ, Audretsch DB, Braunerhjelm P, Carlsson B (2005) Growth and Entrepreneurship: An Empirical Assessment. Papers on entrepreneurship, growth and public policy No. 3205.

Adkins LC, Hill RC (2007) Bootstrap inferences in heteroscedastic sample selection models: A Monte Carlo investigation. Economic Working Papers OKSWP0710.

Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. *Neural Computation* 9: 1545-1588.

Amorós JE, Etchebarne S, Felzensztein C (2012) International entrepreneurship in Latin America: Development Challenges. *ESIC Market Economics and Business Journal* 43: 497-512.

Arenius P, Minniti M (2005) Perceptual variables and nascent entrepreneurship. *Small Business Economics* 24: 233-247.

Austin PC, Tu JV (2004) Bootstrap Methods for Developing Predictive Models. *The American Statiscian* 58: 131-137.

Berrios-Lugo JE, Espina MI (2014) Determinant factos for the development of entrepreneurial activity: A correlational study. *ESIC Market* 147.

Bickel PJ, Ritov Y, Stoker TM (2006) Tailor-Made Tests for Goodness of Fit to Semiparametric Hypotheses. *The Annals of Statistics* 34: 721-741.

Blanchflower DG (2000) Self-employment in OECD countries. *Labour Economics* 7: 471-505.

Blumberg B, Pfann G (2015) Roads leading to self-employment: comparing transgenerational entrepreneurs and self-made starts-ups. IZA DP 9155.

Bosma N, van Praag M, Thurik R, de Wit G (2004) The value of human and social capital investments for the business performance of start-ups. *Small Business Economics* 23: 227-236.

Breiman L (1996) Bagging predictors. *Machine Learning* 24: 123–140.

Breiman L (2001a) Random Forests. *Machine Learning* 45: 5–32.

Breiman L (2001b) Statistical modeling: The two cultures (with comments and re-joinder by the author). *Statistical Science* 16: 199-231.

Brown TE, Ulijn JM (2004) *Innovation, entrepreneurship and culture. The interaction between technology, progress and economic growth*. Northampton, MA: Edward Elgar Publishing.

Büchlmann P, Yu B (2002) Analyzing bagging. *The Annals of Statistics* 30: 927-961.

Coduras A, Clemente JA, Ruiz J (2015) A novel application of fuzzi-set Qualitative Comparative Analysis to GEM Data. *Journal of Business Research* 69: 1265-1270.

Cooper AC, Yin X (2005) Entrepreneurial networks, in *The Blackwell encyclopedia of management –entrepreneurship*, Hitt MA, Ireland RD (eds.). Malden, MA: Blackwell, 98-100.

Davidson R, MacKinnon JG (2006) Bootstrap methods in econometrics. *Mimeo*.

Davidsson P (1989) Entrepreneurship –and after? A study of growth willingness in small firms. *Journal of Business Venturing* 4: 211-226.

Diebold FX, Chen C (1996) Testing structural stability with endogenous breakpoint a size comparison of analytic and bootstrap procedures. *Journal of Econometrics* 70: 221-241.

Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19: 1090-1099.

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7: 1–26.

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

Efron B, Gong G (1983) A leisurely look at the bootstrap the jackknife and cross-validation. *American Statistician* 37: 36-48.

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall.

Freedman DA, Navidi WC (1986). Models for adjusting census. *Statistical Science* 1: 3-11.

Freedman DA, Peters SC (1984) Bootstrapping an econometric model: Some empirical results. *Journal of Business & Economic Statistics*2: 150-158.

Friedman M (1953) The methodology of positive economics.

Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

George EI (2000) The variable selection problem. *Journal of the American Statistical Association* 95: 1304-1308.

Gilbert BA, McDougall PP, Audretsch DB (2006) New venture growth: A review and extension. *Journal of Management* 32: 926-950.

Gong G (1986) Cross-validation, the jackknife and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association* 11: 361-368.

Grimm M, Paffhausen AL (2015) Do Interventions Targeted at Micro-Entrepreneurs and Small and Medium- Sized Firms Create Jobs? A Systematic Review of the Evidence for Low and Middle Income Countries. *Labour Economics* 12: 67–85.

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research* 3: 1157-1182.

Ho TK (1998) The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions* 20: 838-844.

Holcomb TR, Ireland RD, Holmes RM, Hitt MA (2009) Architecture of entrepreneurial learning: exploring the link among heuristics, knowledge, and action. *Entrepreneurship Theory and Practice* 33: 167-192.

Horowitz JL (1997) Advances in economics and econometrics: theory and applications, chapter 7: Bootstrap methods in econometrics: theory and numerical performance. *Econometric Society Monographs* 28: 188-222.

Horowitz JL (2003) The bootstrap in econometrics. *Statistical Science* 18: 211-218.

James G, Witten D, Hastie T, Tibshirani T (2013) *An introduction to Statistical Learning* (Vol. 112). Springer, New York.

Jeong J, Maddala GS (1993) A perspective on application of bootstrap methods in econometrics. *Handbook of Statistics* 11: 573-610.

Karlan D, Valdivia M (2011) Teaching entrepreneurship: Impact of business training on microfinance clients and institutions. *Review of Economics and Statistics* 93: 510-552.

Kelley D (2009) Growth aspirations as a function of entrepreneurial motivations and perceptions. Babson Faculty Research Working Papers 49.

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14); 1137-1145.

Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artificial intelligence* 97: 273-324.

Kotsova T (1997) Country institutional profiles concept and measurement. *Academy of Management Proceedings* 97: 180-184.

Kyrö P (2015) The conceptual contribution of education to research on entrepreneurship education. *Entrepreneurship and Regional Development*: 1-20.

Levie J, Autio E (2013) Growth and growth intentions: A meta-analysis of existing evidence. *Enterprise Research Centre, ERC White Papers* 1.

Li H, Maddala GS (1997) Bootstrapping cointegrating regressions. *Journal of Econometrics*80: 297-318.

Lundstrom A, Stevenson L (2002) On the road to entrepreneurship policy. In *The Entrepreneurship Policy for the Future* (Vol. 1). Stockholm: Swedish Foundation for Small Business Research.

MacKinnon JG (2002) Bootstrap Inference in Econometrics. *The Canadian Journal of Economics* 35: 615-645.

MacKinnon JG (2006) Bootstrap Methods in Econometrics. *The Economic Record* 82: S2-S18.

Minniti M (2005) Entrepreneurship and network externalities. *Journal of Economic Behaviour& Organization* 57: 1-27.

Minniti M, Nardone C (2007) Being in someone else's shoes: Gender and nascent entrepreneurship. *Small Business Economics* 28: 223-239.

Molina JA, Barrado B (2015) Factores macroeconómicos que estimulan el emprendimiento. Un análisis para los países desarrollados y no desarrollados. DTECONZ 2015-06.

Molina JA, Velilla J, Ortega R (2015) The decision to become an entrepreneur in Spain: The role of the household financial situation. *MPRA Papers* 68101.

Murphy KM, Topel RH (2002) Estimation and Inference in Two-Step Econometric Models. *Journal of Business and Economic Statistics* 20: 88-97.

Nagler P, Naudé W (2014) Non-farm enterprises in rural Africa: New empirical evidence. *Policy Research Working Paper* 7066.

Naudé W (2016) Is European Entrepreneurship in Crisis? IZA DP 9817.

Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*: 221-247.

Prasad AM, Iverson LR, Liaw A (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* 9: 181-199.

Quinlan JR (1996) Bagging, boosting, and C4. 5. In *AAAI/IAAI* (Vol. 1); 725-730.

Rao JS, Tibshirani R (1997) The out-of-bootstrap method for model averaging and selection. *University of Toronto.*

Schott T, Bager T (2004) Growth expectations by entrepreneurs in nascent firms, baby business and mature firms. In *The growth of Danish firms* (*Part 2 of the Global Entrepreneurship Monitor*), Bager T, Hancock M (eds.). Copenhagen, DK: BorsensForlag; 219-230.

Schumpeter A (1934) *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*: 8-25.

Stuetzer M, Goethmer M, Cantner U (2012) Do Balanced Skills Help Nascent Entrepreneurs to MakeProgress in the Venture Creation Process? *Economics Letters* 117: 186–188.

Terjesen S, Szerb L (2008) Dice thrown from the beginning? An empirical investigation of firm level growth expectations. *Estudios de Economía* 35: 157-178.

Thurik AR (2009) Entreprenomics: Entrepreneurship, Economic Growth and Policy. *Entrepreneurship, Growth and Public Policy* 219–249. Cambridge: Cambridge University Press.

Veall MR (1992) Bootstrapping the process of model selection: An econometric example. *Journal of Applied Econometrics* 7: 93-99.

Velilla J, Ortega R (2017) Determinants of entrepreneurship using fizzy set methods: Europe vs. non-Europe. *Applied Economics Letters* 24: 1320-1326.

Vinod HD (1993) Bootstrap methods: Applications in econometrics. *Handbook of Statistics* 11: 629-661.

Zheng X, Loh WY (1995) Consistent Variable Selection in Linear Models. *Journal of the American Statistical Association* 90: 151-156.

Figure 1. Pseudo-code

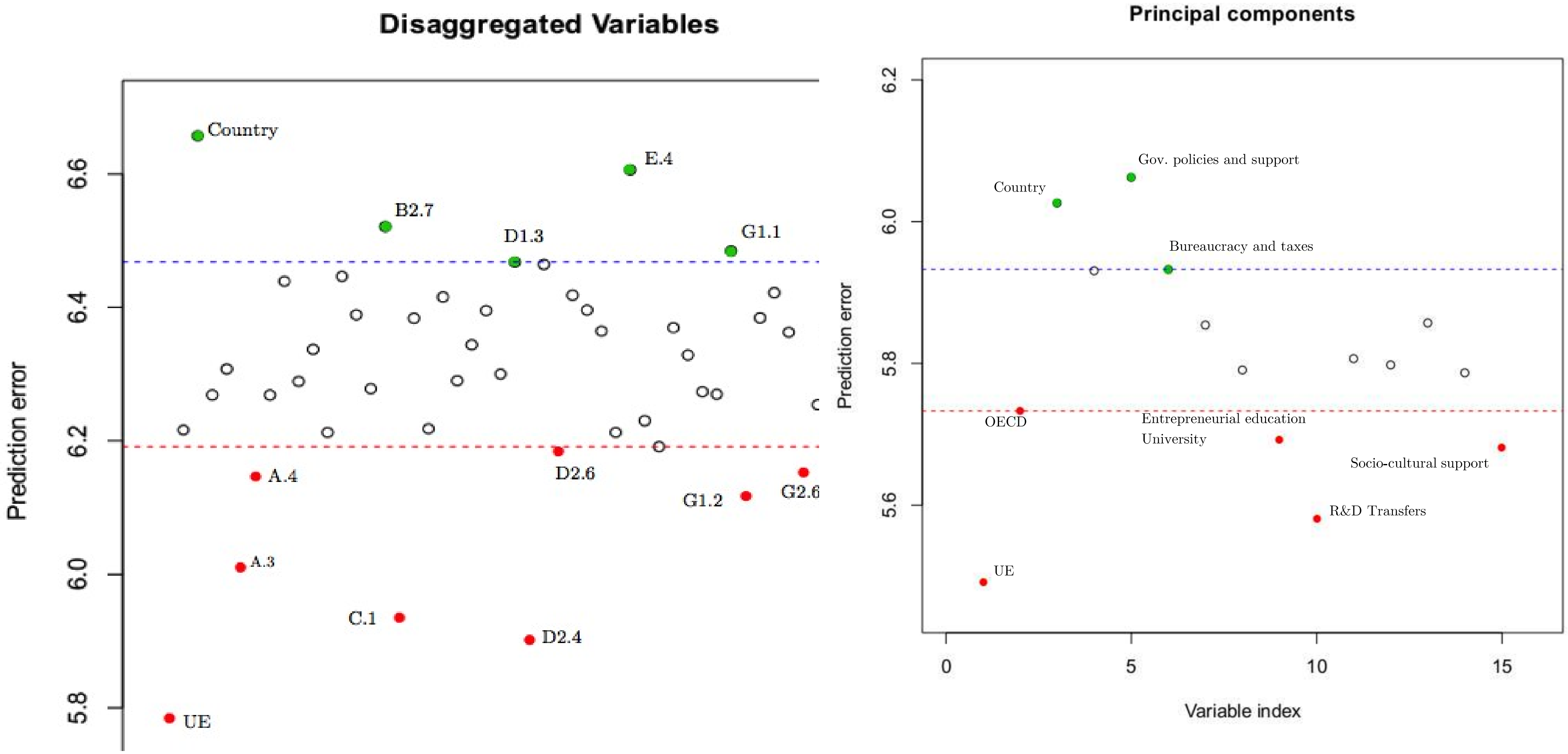
```
Main model:
1. Bootstrap {
   a. Draw a random subsample of observations with replacement of size N.
   b. Draw a random subsample of explanatory variables without replacement of size $m_2$.
   c. Save the indices of the explanatory variables selected in (b).
   d. Estimate the econometric model using individuals selected in (a) and features selected in (b).
   e. Using the unselected individuals of (a), compute the average error of prediction of the model estimated in (d). }
2. For each explanatory variable $X_j$, define its associated m.a.e. using all the bootstrap iterations in which $X_j$ is selected in (b).


Step-by-step model:
1. For each explanatory variable, estimate a bootstrapped model and the corresponding m.a.e. over test sets.
2. Select the variable with the lowest m.a.e.
3. Repeat, including the previously-selected variable in the model and excluding it from (1).
```

Figure 2. Results for GEM 2014 NES National Level data

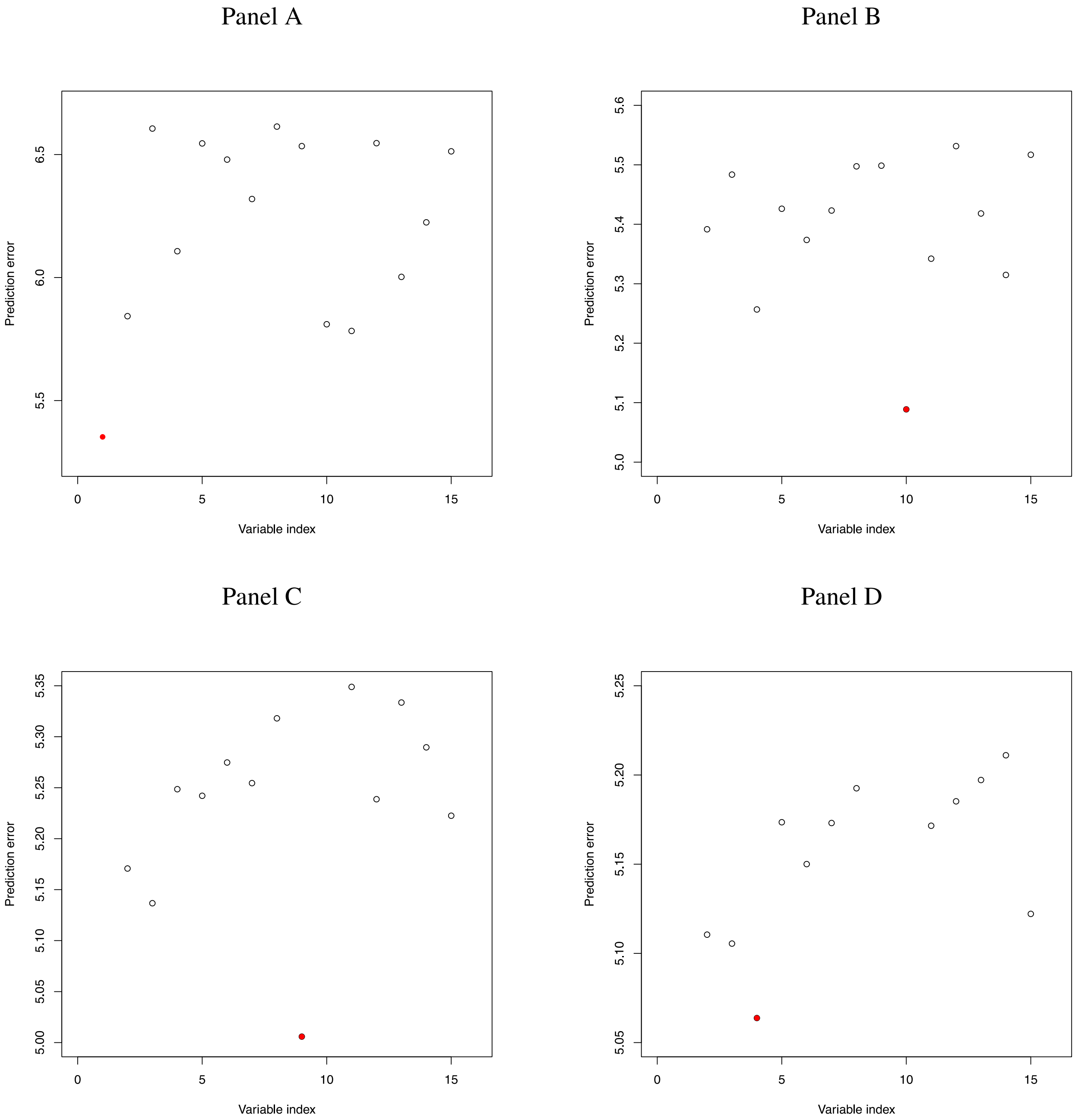
Panel A

Panel B



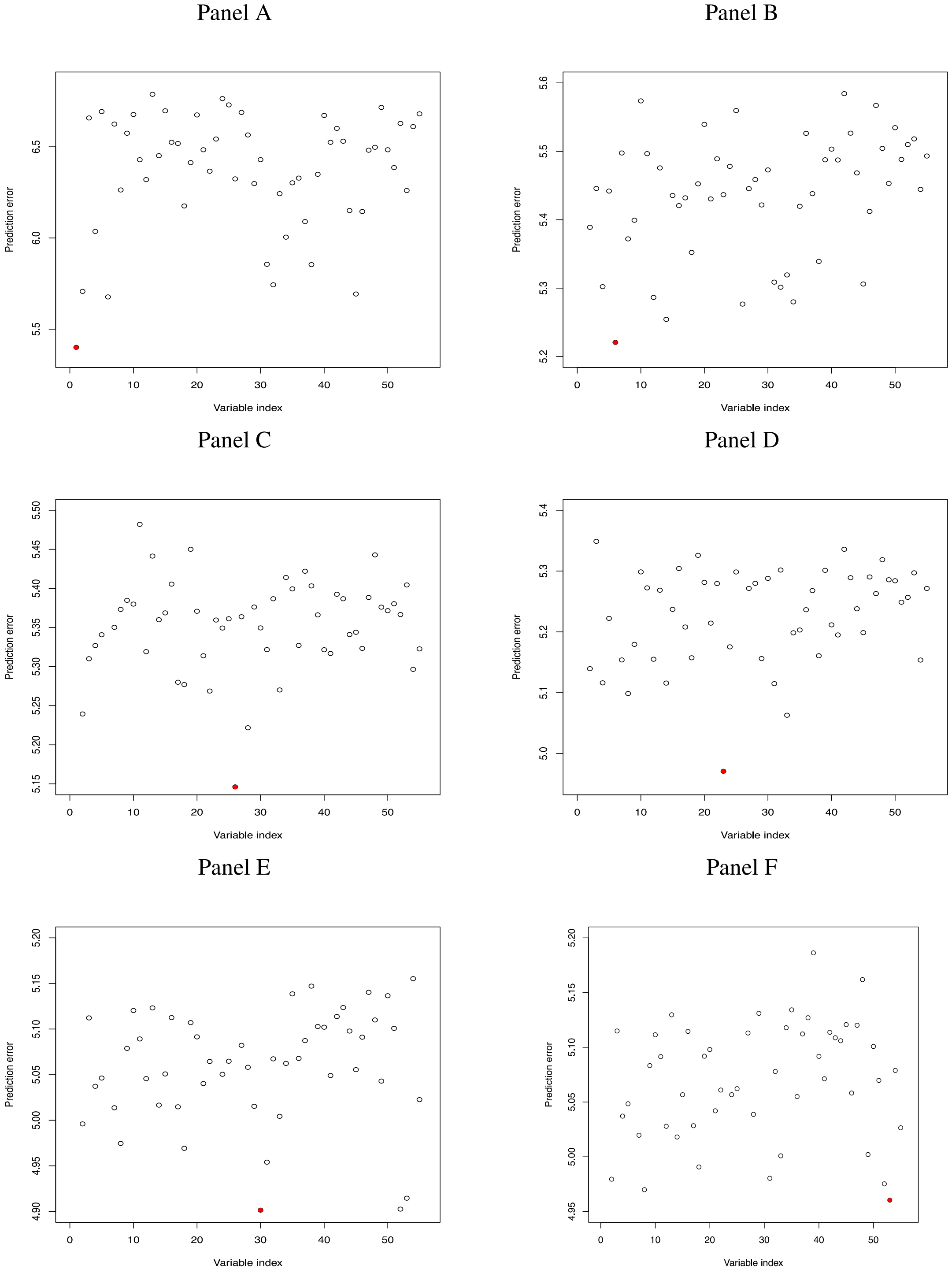
Note: the sample is from GEM 2014 NES National. Dependent variable is the "TEA Index". In Panel A, we use the disaggregated variables, while in Panel B we use aggregated principal component factors (variable indices are taken ordered from Table 1). We compute 500 iterations bootstrap of linear regressions, following the method described in Section 2.

Figure 3. Forward stepwise for principal factors analysis of GEM 2014 NES National data



Note: the sample (GEM 2014 NES National) is restricted to individuals who reported age. Dependent variable is the "TEA index". Regressors are aggregated principal component factors (see Table 1). In Panel A, we show results of 2,000 iterations bootstrap of linear regressions with a single regressor in each iteration. The variable with the highest predictive power is "EU" (index 1). In Panel B, we repeat the process selecting "EU" and another variable in each iteration. The variable with the highest predictive power is now the R&D transfers level (index 10). Repeating analogously, in Panel C, the variable with the highest predictive power is now the level of vocational, professional, College, and University entrepreneurial education (index 9). Finally, in panel D, the selected variable is the financial environment (index 4). Prediction errors (mean absolute errors) considerably decrease in Panel A, B and C, while they increase in Panel D.

Figure 4. Forward stepwise for all individual variables of GEM 2014 NES National data



Panel A

Panel B

Panel C

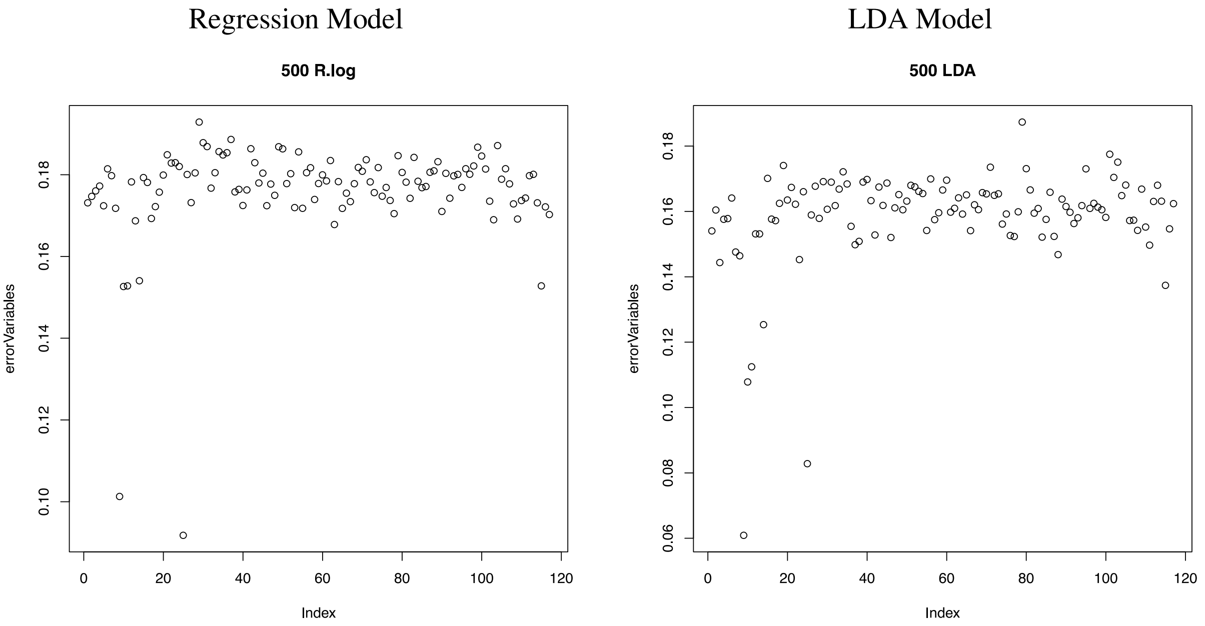
Panel D

Panel E

Panel F

Note: the sample (GEM 2014 NES National) is restricted to individuals who reported age. Dependent variable is "TEA index". Regressors are disaggregated variables (see Table 1). In Panel A, we show results of 2,000 iterations bootstrap of linear regressions with a single regressor in each iteration. The variable with the highest predictive power is "EU" (index 1). In Panel B, we repeat the process selecting "EU" and another variable in each iteration. The variable with the highest predictive power is now government subsidies for new and growing firms (index 6). Repeating analogously, in Panel C, the variable with the highest predictive power is now entrepreneurial preparation at University (index 26); in Panel D, creativity, self-sufficiency, and initiative at primary/secondary school (index 23); and in Panel E, equity of access to technology by new and growing forms (index 30). Finally, we repeat the process in panel F and the selected variable is now the culture of encouragement of entrepreneurial risk-taking (index 53). Prediction errors (mean absolute errors) considerably decrease in Panel A, B, C, D and E but increase in Panel F.

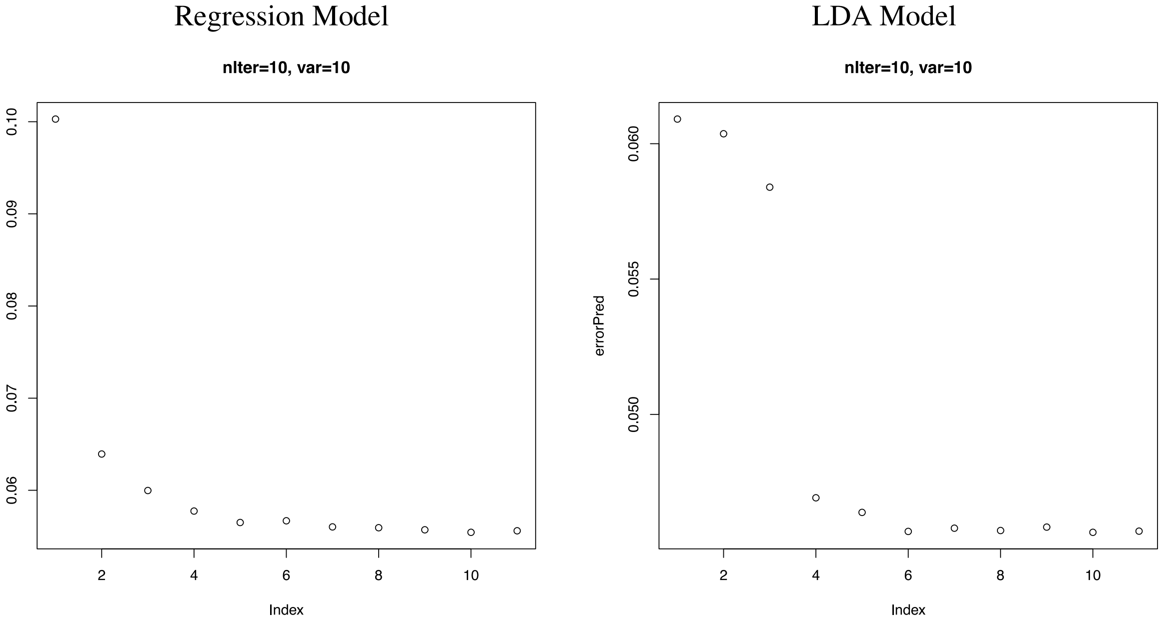
Figure 5. Application of the algorithm to classification problems with GEM 2014 APS Global data


Regression Model

LDA Model

Note: the sample (GEM 2014 APS Global) is restricted to individuals who reported age. Dependent variable is the dummy "contributes to TEA". Panel A shows results for a logistic regression problem, and Panel B for a Linear Discriminant Analysis. We compute 500 iterations bootstrap of logistic and LDA classification problems, following the method described in Sections 2.

Figure 6. Forward stepwise for classification problems with GEM 2014 APS Global data



Note: the sample (GEM 2014 APS Global) is restricted to individuals who reported age. Dependent variable is the dummy "contributes to TEA". In Panel A, we show results of 10 iterations bootstrap of logistic regressions following a forward stepwise procedure. In Panel B, we repeat the process for an LDA.

Table 1. Macro-economic factors- GEM 2014 NES National

		Mean	S. Dev
TEA	*Total* (*Early-Stage*)Entrepreneurial *Activity index*	13.084	8.156
Country	*Country name/code*	-	-
EU	*Belonging to EU*	0.333	0.474
OECD	*Belonging to the OECD*	0.420	0.497
Group A	(*Financial environment related to entrepreneurship*)	2.535	0.397
A.1	*In my country, there is sufficient equity funding available for new and growing firms*	2.595	0.499
A.2	*In my country, there is sufficient debt funding available for new and growing firms*	2.684	0.434
A.3	*In my country, there are sufficient government subsidies available for new and growing firms*	2.715	0.588
A.4	*In my country, there is sufficient funding available from private individuals (other than founders) for new and growing firms*	2.574	0.388
A.5	*In my country, there is sufficient venture capital funding available for new and growing firms)*	2.460	0.480
A.6	*In my country, there is sufficient funding available through initial public offerings (IPOs) for new and growing firms*	2.286	0.590
Group B1	(*Government concrete policies, priority, and support*)	2.617	0.427
B1.1	*In my country, Government policies (e g, public procurement) consistently favor new firms*	2.237	0.407
B1.2	*In my country, the support for new and growing firms is a high priority for policy at the national government level*	2.875	0.532
B1.3	*In my country, the support for new and growing firms is a high priority for policy at the local government level*	2.716	0.492
B1.4	*In my country, new firms can get most of the required permits and licenses in about a week*	2.227	0.677
Group B2	(*Government policies bureaucracy, taxes*)	2.462	0.569
B2.5	*In my country, the amount of taxes is NOT a burden for new and growing firms*	2.455	0.616
B2.6	*In my country, taxes and other government regulations are applied to new and growing firms in a predictable and consistent way*	2.767	0.642
B2.7	*In my country, coping with government bureaucracy, regulations, and licensing requirements is not unduly difficult for new and growing firms*	2.432	0.611
Group C	(*Government programs*)	2.679	0.438
C.1	*In my country, a wide range of government assistance for new and growing firms can be obtained through contact with a single agency*	2.423	0.598
C.2	*In my country, science parks and business incubators provide effective support for new and growing firms*	3.106	0.571
C.3	*In my country, there are an adequate number of government programs for new and growing businesses*	2.896	0.559
C.4	*In my country, the people working for government agencies are competent and effective in supporting new and growing firms*	2.706	0.428
C.5	*In my country, almost anyone who needs help from a government program for a new or growing business can find what they need*	2.448	0.447
C.6	*In my country, Government programs aimed at supporting new and growing firms are effective*	2.625	0.414
Group D1	(*Entrepreneurial level of education at Primary and Secondary*)	2.091	0.398
D1.1	*In my country, teaching in primary and secondary education encourages creativity, self-sufficiency, and personal initiative*	2.270	0.486
D1.2	*In my country, teaching in primary and secondary education provides adequate instruction in market economic principles*	2.091	0.410
D1.3	*In my country, teaching in primary and secondary education provides adequate attention to entrepreneurship and new firm creation*	1.944	0.390
Group D2	(*Entrepreneurial level of education at vocational, professional, College and University*)	2.891	0.306
D2.4	*In my country, Colleges and universities provide good and adequate preparation for starting up and growing new firms*	2.694	0.342
D2.5	*In my country, the level of business and management education provides good and adequate preparation for starting up and growing new firms*	3.080	0.312
D2.6	*In my country, the vocational, professional, and continuing education systems provide good and adequate preparation for starting up and growing new firms*	2.936	0.361
Group E	(*R&D level of transference*)	2.393	0.379
E.1	*In my country, new technology, science, and other knowledge are efficiently transferred from*	2.360	0.374

	universities and public research centers to new and growing firms		
E.2	*In my country, new and growing firms have just as much access to new research and technology as large, established firms*	2.297	0.367
E.3	*In my country, new and growing firms can afford the latest technology*	2.142	0.356
E.4	*In my country, there are adequate government subsidies for new and growing firms to acquire new technology*	2.347	0.527
E.5	*In my country, the science and technology base efficiently supports the creation of world-class, new technology-based ventures in at least one area*	2.831	0.579
E.6	*In my country, there is good support available for engineers and scientists to have their ideas commercialized through new and growing firms*	2.528	0.528
Group F	(*Professional and commercial infrastructure access*)	3.024	0.333
F.1	*In my country, there are enough subcontractors, suppliers, and consultants to support new and growing firms*	3.322	0.428
F.2	*In my country, new and growing firms can afford the cost of using subcontractors, suppliers, and consultants*	2.418	0.325
F.3	*In my country, it is easy for new and growing firms to get good subcontractors, suppliers, and consultants*	2.813	0.390
F.4	*In my country, it is easy for new and growing firms to get good, professional legal and accounting services*	3.349	0.443
F.5	*In my country, it is easy for new and growing firms to get good banking services (checking accounts, foreign exchange transactions, letters of credit, and the like)*	3.314	0.486
Group G1	(*Internal market dynamics*)	2.968	0.475
G1.1	*In my country, the markets for consumer goods and services change dramatically from year to year*	3.021	0.529
G1.2	*In my country, the markets for business-to-business goods and services change dramatically from year to year*	2.907	0.446
Group G2	(*Internal market burdens*)	2.630	0.337
G2.3	*In my country, new and growing firms can easily enter new markets*	2.730	0.331
G2.4	*In my country, the new and growing firms can afford the cost of market entry*	2.517	0.318
G2.5	*In my country, new and growing firms can enter markets without being unfairly blocked by established firms*	2.634	0.360
G2.6	*In my country, the anti-trust legislation is effective and well-enforced*	2.689	0.508
Group H	(*Physical infrastructure and services access*)	3.735	0.483
H.1	*In my country, the physical infrastructure (roads, utilities, communications, waste disposal) provides good support for new and growing firms*	3.421	0.678
H.2	*In my country, it is not too expensive for a new or growing firm to get good access to communications (phone, Internet, etc)*	3.836	0.494
H.3	*In my country, a new or growing firm can get good access to communications (telephone, internet, etc) in about a week*	3.899	0.523
H.4	*In my country, new and growing firms can afford the cost of basic utilities (gas, water, electricity, sewer)*	3.775	0.483
H.5	*In my country, new or growing firms can get good access to utilities (gas, water, electricity, sewer) in about a month*	3.820	0.516
Group I	(*Cultural, social norms and society support*)	2.848	0.398
I.1	*In my country, the national culture is highly supportive of individual success achieved through own personal efforts*	3.043	0.493
I.2	*In my country, the national culture emphasizes self-sufficiency, autonomy, and personal initiative*	2.901	0.464
I.3	*In my country, the national culture encourages entrepreneurial risk-taking*	2.507	0.440
I.4	*In my country, the national culture encourages creativity and innovation*	2.891	0.391
I.5	*In my country, the national culture emphasizes the responsibility that the individual (rather than the collective) has in managing his or her own life*	2.894	0.412

Note: The sample is GEM 2014 NES National Level (69 observations). "EU" and "OECD" are *dummy* variables. TEA takes value in percentage. The rest of the variables take values between 0 (disagree) and 5 (agree). For each country in the sample, group values are the mean of the single attributes that belong to the group.

Table 2. Summary statistics - GEM 2014 APS Global

	OECD		Non-OECD		
Variables	Mean	S. Dev.	Mean	S. Dev.	*p*-value
Contributes to TEA index	0.0813	0.2733	0.1621	0.3686	(<0.01)
Age	43.043	14.700	38.158	13.656	(<0.01)
Being male	0.4895	0.4998	0.4788	0.4995	(<0.01)
Family number	3.2051	1.6032	4.1119	2.1926	(<0.01)
Number of children	0.8238	1.3321	0.2612	0.8671	(<0.01)
Basic education	0.1858	0.3889	0.0499	0.2178	(<0.01)
Secondary education	0.5879	0.4922	0.6248	0.4841	(<0.01)
University education	0.1639	0.3702	0.2929	0.4551	(<0.01)
Consider to have skills to be entrepreneur	0.4396	0.4963	0.5482	0.4976	(<0.01)
Low level of income	0.2554	0.4361	0.3021	0.4592	(<0.01)
Middle level of income	0.2503	0.4332	0.2682	0.4432	(<0.01)
High level of income	0.2470	0.4313	0.2672	0.4425	(<0.01)
Being employed	0.5460	0.4978	0.3944	0.4887	(<0.01)
Being self-employed	0.1411	0.3471	0.2895	0.4887	(<0.01)
Being unemployed	0.1170	0.3214	0.1353	0.3420	(<0.01)
Being an active unemployed	0.1166	0.3209	0.1349	0.3416	(<0.01)
Being retired	0.1336	0.3402	0.0885	0.2688	(<0.01)
Being a student	0.1024	0.3019	0.1154	0.3195	(<0.01)
Being a homemaker	0.1338	0.3404	0.2510	0.4335	(<0.01)
Being a businessman	0.1069	0.3088	0.1927	0.3944	(<0.01)
Desire to be a businessman	0.4396	0.4963	0.6618	0.4730	(<0.01)
Desire to be entrepreneur in the future	0.1393	0.3463	0.3200	0.4665	(<0.01)
Know someone who is an entrepreneur	0.3916	0.8981	0.5032	0.9155	(<0.01)
Consider to have opportunities to entr.	0.2684	0.4431	0.4364	0.4959	(<0.01)
Have a good perception of entrepreneur	0.2682	0.4430	0.4406	0.4964	(<0.01)
Consider that culture supports entrep.	0.4281	0.4948	0.6671	0.4712	(<0.01)
Have fear of failure	0.4385	0.4962	0.3543	0.4783	(<0.01)
Desire for social equity	0.4931	0.4999	0.5670	0.4954	(<0.01)
Respect for success	0.5412	0.4983	0.6673	0.4711	(<0.01)
Media promotes successful businessmen	0.4155	0.4928	0.6420	0.4794	(<0.01)
Have helped other entrepreneurs	0.0707	0.2564	0.1262	0.3321	(<0.01)
Can offer a new product	0.0637	0.2442	0.0950	0.2932	(<0.01)
Can offer a new technology	0.0344	0.1823	0.0750	0.2634	(<0.01)
Technological level of the sector working	0.0096	0.0975	0.0047	0.0685	(0.0651)
Search for new markets	0.1503	0.3574	0.2619	0.4397	(<0.01)
Search for new customers	0.0031	0.0562	0.0158	0.1248	(<0.01)
Have ended a business recently	0.0288	0.1674	0.0609	0.2392	(<0.01)
Have invested in other business	0.0487	0.2155	0.0599	0.2374	(<0.01)
Belong to a developing country	-	-	0.7721	0.4194	(<0.01)
Belong to a non-developed country	-	-	0.1680	0.3739	(<0.01)
Have an ideal life	0.4969	0.4999	0.5288	0.4991	(<0.01)
Have good conditions of life	0.4986	0.5000	0.5022	0.4999	(0.1153)
Be satisfied with life	0.6782	0.4671	0.6177	0.4859	(<0.01)
Obtain important things in life	0.6028	0.4893	0.5173	0.4997	(<0.01)
Satisfied with life decisions	0.4409	0.4965	0.4257	0.4944	(<0.01)
Have freedom at work	0.3452	0.4754	0.3307	0.4704	(<0.01)
Like current job	0.8932	0.3088	0.8849	0.3201	(<0.01)
Have a non--stressful job	0.6852	0.4644	0.7770	0.4162	(<0.01)
Satisfied with current job	0.3788	0.4851	0.3018	0.4590	(<0.01)
Satisfied with current income	0.2322	0.4222	0.2226	0.4160	(<0.01)
Enough time for leisure	0.1947	0.3959	0.2965	0.4567	(<0.01)
Enough time for family activities	0.2007	0.4005	0.3097	0.4623	(<0.01)
Enough time for housework activities	0.2187	0.4133	0.3165	0.4651	(<0.01)
Being an immigrant	0.0487	0.2153	0.0113	0.1060	(<0.01)
N. Observations	93,635		94,738		

Note: the sample (GEM 2014 APS Global) is restricted to individuals who reported age. Kruskal-Wallis *p*-values for the comparison of the variables among belonging or not belonging to the OECD in parenthesis. Age is measured in years. The rest of the variables are dummies, taking value 1 if agree or 0 if disagree.

Table 3. Measures of goodness of fit. Regression models

	Principal factors			Disaggregated variables		
MEASURES	**5.1 Model**	**5.2 Model**	**Stepwise selection**	**5.1 Model**	**5.2 Model**	**Stepwise selection**
AIC	469.333	466.150	463.495	470.447	465.213	456.122
BIC	482.738	475.087	474.666	495.022	476.383	467.292
R^2	0.325	0.317	0.361	0.407	0.345	0.426
Adjusted R^2	0.271	0.285	0.321	0.304	0.304	0.390
Prediction error	5.223	5.030	5.308	5.013	4.928	5.070

Note: the sample is GEM 2014 NES National data. Dependent variable is the TEA index. Prediction errors, defined as the absolute average difference between predicted and current values of the dependent variable, are computed over test sets.

Table 4. Measures of goodness of fit. Logistic models

MEASURES	6.1 Model	6.2 Model	Stepwise selection
AIC	57139.66	59387.47	51412.07
BIC	57200.51	59488.93	53159.17
Pseudo R^2	0.589	0.575	0.621
Prediction error	0.061	0.057	0.081

Note: the sample is GEM 2014 APS Global, restricted to individuals who reported age. Dependent variable is the dummy "contributes to TEA". Prediction errors, defined as the absolute average difference between predicted and current values of the dependent variable, are computed over test sets.

Table 5. Final logit model

VARIABLES	Estimate	Std. Error	*t*-ratio	***p*-value**
Have helped other entrepreneurs	2.830***	0.039	72.225	(<0.001)
Age	-0.045***	0.001	-39.204	(<0.001)
Search for new markets	22.669	72.700	0.312	(0.755)
Being a businessman	-1.506***	0.043	-34.746	(<0.001)
Have ended a business recently	0.134*	0.043	2.471	(0.013)
Constant	-20.077	72.700	-0.276	(0.782)
N. Obs.	188,373			

Note: the sample (GEM 2014 APS Global) is restricted to individuals who reported age. Dependent variable is the dummy "contributes to TEA".