Wubetie *J Big Data* (2019) 6:17
https://doi.org/10.1186/s40537-018-0153-4

Journal of Big Data

METHODOLOGY Open Access

Application of variable selection and dimension reduction on predictors of MSE's development

Habtamu Tilaye Wubetie*

*Correspondence:
habtamu.tilaye@yahoo.com
Statistics Department, College of Natural and Computational Science, University of Gondar, Gondar, Ethiopia

Abstract

Nature create variables using its character component, and variables are sharing characters from a vary small to relatively large scale. This results, variables to have from a vary different to a more similar character, and leads to have a relation ship. Literature suggested different relation measures based on the nature of variable and type of relation ship exist. Today, due to having high variety of frequently produced large data size, currently suggested variable filtering and selection methods have gaps to full fill the need. This research desires to fill this gap by comparing literature suggested methods to finding out a better variable selection and dimension reduction methods. The result from regression analysis using all literature suggested factors shows that none of the predictors for development status of enterprise are significant, and only 10 predictors for number of employer in an enterprise are significant out of 81 factors. Since, variable selection and dimension reduction methods are applied to find out predictors of a response by removing variable redundancy, and complexity of incorporating large number variable. Based on statistical power, for the results from variable selection methods, specially association and correlation methods showed that, CANOVA more efficiently detects non-linear or non-monotonic correlation between a continuous–continuous and a continuous-categorical variables. Spearman's correlation coefficient more efficiently detects a monotonic correlation between a continuous with a continuous, and a continuous with a categorical variable. Pearson correlation coefficient more efficiently detects the linear correlation between continuous variables. MIC efficiently detects non-linear or non-monotonic relation between continuous variables. Chi-square test of independence efficiently detects relation between a continuous with a continuous, and categorical with categorical variables, but the non linear or non monotonic relation between a continuous with a categorical are not well detected. On the other hand, the result from lasso and stepwise methods reveals that, the relation between the predictor and response due to interaction effect not detected by correlation and association methods are detected by stepwise variable selection method, and the multicollinearity is detected and removed by lasso method. Regressing the response variable "number of employer in an enterprise" based on variables selected by lasso and stepwise method does bring greater model fitness (based on adjusted R-squared value) than variables selected by association and correlation methods. Similarly, regressing the response variable "development status of an enterprise" based on variables selected by association and correlation methods does bring

Springer Open

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

12 significant variables, where none of variables are significant from variables selected by lasso and stepwise methods. As a result, 51 predictors for number of employment in an enterprise, and 40 predictors for development status of an enterprise are detected as significantly related variables. And, lasso and stepwise methods are preferred to select predictors of a continuous response variable "number of employers in an enterprise", and association and correlation methods are preferred to select predictors of a categorical response variable "development status of an enterprise". Finally, the reduced regression models result reveals that, 20 predictors have causal relation with number of employment in an enterprise, and 12 predictors have causal relation with development status of an enterprise. On the other hand, based on model fitness, information lost, and number of significant factors, principal factor is preferred and applied in dimension reduction for a categorical response variable "development status of an enterprise", and factor score based regression is preferred and applied for a continuous response variable "number of employers in an enterprise". However, the comparison of the results in variable selection and dimension reduction indicates that, variable selection methods gave more gain in model fitness than dimension reduction methods. Hence, the suggested variable selection methods are more preferred than dimension reduction methods, and applied to find out predictors. In general, the suggested procedure for variable selection methods are recommended when small number of variables are studied, and the suggested dimension reduction methods are recommended for large number of variant variables (Big data case).

Keywords: Variable selection, Dimension reduction, CANOVA, Stepwise elimination, Lasso variable selection

Introduction

Nature create variables using its character component, and variables are sharing characters from a vary small to relatively large scale. This results, variables to have from a vary different to a more similar character. Variables having a more similar character are variables sharing largely a more similar character component (have relatively the same composition), and apparently a vary small similarity is due to high difference in component character composition. Hence, taking variables having more similar character as one variable or taking one of them as a representative can remove natural character redundancy, and it helps to mange and analyse the relation ship between variables in a world of large amount of variables are inter-related. This inter-relation between the variables causes the variables to have a direct causal relation, or an indirect causal relation or relation with out causal nature. Statistically, a direct causal relation indicates the presence of dependency between variables, where as indirect causality is due to the presence of latent variable. However, the relation between the variables without known causality is due to not well understood relation in the real world. The relation between variables can be linear or non-linear or random. Statistical methods like, variable-selection and variable-dimension-reduction methods can used to reduce the number of variable by taking single variable or merging as a component for statistically significantly similar variables.

Measuring the predictor–predictor relation, and response–predictor relation is important to recognize the relationship exist, and having a short list of influential factors for further analysis to determine their effect on response variable.

However, due to inter-relation between dependent variables, their influence on response variable is not only individual rather in group too. Since, the natural inter-relation between variable is not captured and considered by simulation study, or by predictor–response

association or correlation measures only. Correspondingly, this interaction effect is planed to detected for real data using Micro and small enterprise (MSE's) data set[File Name: MSEs.csv] by considering the predictors filtered by association, correlation and regression measures for predictor–predictor and predictor–response relation. Then, the possible combination of selected (filtered) groups of variables are then regressed for response variable, and significantly and potentially related variables are re-selected using stepwise and lasso variable selection method.

Statistical measures of association, correlations and regression are used to find out the relation exist between variables. In this research the statistical relation measures used for variable selection, and dimension reduction are, Pearson correlation coefficient, Spearman's rank correlation coefficient, Chi-square test of independence, maximal information criterion (MIC), continuous analysis of variance test (CANOVA), stepwise variable selection and lasso variable selection, and Principal factor and Factor score analysis respectively.

Wang et al. [20] used simulated and real datasets (kidney cancer RNA-seqdataset) to compare the false positive rates and statistical power of CANOVA to six other methods (Distance correlation's, Hoeffding's independence test, CANOVA the Pearson correlation coefficient, the Spearman's rank correlation coefficient, the Kendall's rank correlation coefficient and the Maximal information coefficient), and showed that CANOVA, the Pearson correlation coefficient, the Spearman's rank correlation coefficient, the Kendall's rank correlation coefficient and the MIC gave the expected false positives. Hence, these methods can detect the true significant variables. However, the false positive rate is lower than the expected for distance correlation and higher than the expected for Hoeffding's independence test. So the true significant variables may not be detected by distance correlation, and there may be false significant variables in Hoeffding's independence test result. Hence, Pearson correlation were recommended when correlation between two continuous variable is linear, and CANOVA were recommended when the correlation between two continuous variable is non-linear or complicated.

Variable dimension reduction is a tool to avoid complexity due to having large number of variables by considering the possible small number of variables those can reflect the needed information: which arise due to some variables are highly correlated to each other or to latent variable, or from the set of variables some variables may accounted for large amount of variability in the data set. For this type of problem variable reduction methods like principal factor analysis and factor score analysis are suggested [1, 2].

Currently due to having high variety of frequently produced big data size, literature suggested variable filtering and selection methods have gaps to full fill the need. Hence, this research desires to fill this gap by finding out a better variable filtering, selection and dimension reduction methods using real data. The above statistical methods of variable-selection and variable-dimension-reduction are applied to reduce the number of variable by taking single variable or merging as a component for statistically significantly similar variables.

Data and variable

From literature, entrepreneur's development is measured in relation to the success of an individual, society, and firm survival [3, 4]. Bosma et al. [4] measured development of enterprise by considering profits of the entrepreneur, employment created by the entrepreneur, and the survival period of the firm. The determinants for development of

entrepreneurs are dependent on the starting human capital, social capital, financial capital and strategies applied on business.

Coduras et al. [5] construct a measure for an individual's readiness for entrepreneurship based on three main categories: sociological, psychological and managerial–entrepreneurial. The South African small enterprise development agency perform a study based on literature and current data for the impact of 2008 and 2009 global financial crisis on South Africa's SMMEs, and they suggests that the South Africa's SMMEs are challenged by access to finance and markets, poor infrastructure, labour laws, crime, skills shortages and inefficient bureaucracy. Assefa et al. [7] perform a study on factors affecting the success of Micro and Small-scale Enterprises in Addis Ababa and five other major regional towns in Ethiopia and find out the key success factors are personal qualities, such as having an articulate vision or ambition and innate abilities, working experience in the formal sector as a factory employee or having worked in family businesses, managerial and entrepreneurial skills, and higher equity in the invested money. Whereas shortage and small size of credit, shortage of working and sales spaces, lack of rental machinery and stringent licensing requirements are constraints of MSEs.

The sample data is taken from Debre Markos town enterprises in 2017. The study units are individuals starting their business in the interval of a year 1994 to 2006 and currently working on their own enterprise or business. The respondents gave detailed information on their entrepreneurial knowledge, skill and experience, on business environment and their strategies. Additional information on enterprises were also taken from Trade and industry office of Debre Markos town.

Sampling method of a study is determined based on the nature of the population under study. Ethiopian Ministry of Urban Development and Housing (MoUDH) classify micro and small size enterprise into five sectors, namely Manufacturing sector, Service sector, Trade, Construction sector, service sector, and Mining and Quarrying Sector. However, based on the present Trade and industry office of Debre Markos town MSEs are re-classified as Manufacturing sector, Service sector, Trade, Urban farming and Construction sector, by splitting Service sector in to service and Urban Farming. Hence, enterprises across sector are more heterogeneous than within sector, stratified sampling method is the right choice. The sample size is determined by using stratified optimal allocation based on the strata's variance calculated from the information (secondary data) obtained from Trade and industry office: for the situation in which the variable of interest is enterprise development status which is categorical with value 1 (achieved expected progress stated by MoUDH) and 0 (not achieved expected progress), and at 99% level of confidence for the true population proportion to be in 0.05 interval of the sample proportion, 179 sample of enterprise is taken from a total of 2093 enterprises. The study unites are allocated to each strata by considering strata's variance rather than proportion, due to high difference in strata's size where some clusters have size less than 20 and some larger than a thousand [8].

Variable of the study

Under these study two dependent and 81 independent variables are considered. List of explanatory variables considered are listed in Appendix: Tables 12, 13, 14, 15, 16 and 17.

Dependent variable

The variable of interest is enterprise development status. Bosma et al. [4] measured Entrepreneurs development (which is individual approach to measure enterprise development status) in relation to, the success of an individual like profit made and capital growth, the success of society based on employee capacity, and firm survival. Contextually, Ethiopian Ministry of Urban Development and Housing (MoUDH) state a measure for development status of micro and small size enterprise based on the progress made by an enterprise on their capital accumulation and human capital mainly in terms of number of employee [3]. The MoUDH definition for micro and small enterprise is given by Table 1.

Correspondingly, on this study enterprise development status is measured based on the progress made by an enterprise which is a categorical variable with value 1 (achieved expected progress) and 0 (not achieved expected progress), and by number of employers in an enterprise as defined by MoUDH.

Explanatory variables

Explanatory variables or factors those have direct or indirect influences on interest variable is the concern need to dig out to find out relevant solution on achieving the planed enterprise development by controlling influential variables. As stated on literature by Bosma et al. [4] determinants for development of entrepreneurs are related to starting human capital, social capital, financial capital, and strategies applied on business.

In general, literature suggested measures of control variables, human capital, financial capital, influencing factors, social capital, and information's relevant for the development of their businesses are considered [3–19].

Variable-selection method

Chi-squared test of independence

Chi-square test of independence is one of the statistical measures that tests the linear and non-linear association between variables. This test helps to determine whether variables are independent of each other or whether there is pattern of dependency between variables. Formally, chi square test of independence determine whether the observed pattern between the variables is strong enough to show that the two variables are dependent on each other, or by considering all possible combinations of variables events and testing for the independence of each pair of these events. If the probability of occurrence of the different possible values of one variable depend on which category of another variable occurs, then the two variables are dependent on each other. Chi-square variable have a continuous distribution obtained by the sum of the squares of a set of

Table 1 Current definition of MSEs in Ethiopia

Level of enterprise	Sector	Head count staff	Total asset ETB	Total asset USD
Micro enterprise	Industry	≤ 5	≤ 100,000	≤ 4630
	Service	≤ 5	≤ 50,000	≤ 2310
Small enterprise	Industry	6–30	101,000–1,500,000	4630–69,500
	Service	6–30	50,001–500,000	2310–23,150

normally distributed variables. Chi-square distribution is a rightly skewed distribution with lower limit at 0 and declines as χ^2 increases to the right with most of values near the center of the distribution. Since, theoretical distribution of chi square distribution is a continuous distribution, and the chi square statistic have discrete distribution, chi square statistic is approximated by the theoretical chi square distribution for reasonably large sample size or for expected number of cases exceed 5 in most cells of the cross classification table. The wildly used rule on expected cases are less than 1 and no more than 20% of expected cases have less than 5 per category. The chi square test for independence is conducted by assuming that there is no relationship (independent) between the two variables being examined versus an alternative hypothesis clam: there is some relationship (dependency) between the variables. Under the null hypothesis of no relationship between variables, the expected cases for each of the cell can be obtained from the multiplication rule of probability for independent events.

Continuous analysis of variance test (CANOVA)

CANOVA is a measure for non linear correlation between two continuous variables, as an extension to ANOVA for continuous variables by making generalization on "within category variance". CANOVA first define a neighborhood for each data point of response variable based on its predictor value, and then the variance of the response value within the neighborhood is calculated. The hypothesis of CANOVA "similar neighbor predictor values lead to similar response values" is tested for smaller value of statistic "within neighborhood sum square" compared to "random expectation". Since, a statistic "within neighborhood variance" does not follow any familiar distribution, its significance is tested by permutation test. The grid of a larger K has more power on slow-varying functions, while a smaller K has more power on quick-oscillating functions depending on the data. The suggested choice for the neighborhood structure of the dataset is $n/20$ [20]. CANOVA is related to local regression (like, K nearest neighbor (kNN) regression), and CANOVA can be viewed as an analogy of the model fitness test of the kNN model as Pearson's correlation coefficient can be viewed as the model fitness test of a linear regression model. This method reduce algorithm complexity to O(nlogn+np) by ordering the data values of response with respect to the ordered value of predictors, and can easily explore the non linear correlation between two continuous variable.

Maximal information criterion (MIC)

MIC is an equitable maximal information-based non-parametric exploration (MINE) statistic for identifying and classifying relationships. This implies, in addition to measuring association, MIC measures non-linear relation between two random variables, and the degree of linear relation between variables having functional relationships. In general, with sufficient sample size it captures all type of functional relationships even that are not well modelled. MIC assigns a score measures strength of relationship in a rage of 0 to 1, where a score of 0 to statistically independent variables, and a score of 1 in probability for noiseless functional relationships. For large data set with many variables (Big data) which contain important and undiscovered relationships, MINE helps in identifying and characterizing structures in data for variable selection or dimension reduction purpose [21].

Pearson correlation coefficient

The Pearson correlation coefficient is the most commonly used correlation method to measure a two-way linear correlation, calculated by dividing covariance of two variables by the product of their standard deviations'. Its value is represented by (r_{xy}) in a range between -1 and 1. If the points (x_i, y_i) are in a perfect straight line and the slope of that line is positive, $(r_{xy}) = 1$. If the points are in a perfect straight line and the slope is negative, $(r_{xy}) = -1$. If there is no systematic relation between X and Y at all, $(r_{xy}) \simeq 0$, and (r_{xy}) differs from zero only because of random variation in the sample points.

Coefficient of determination which is the square of Pearson correlation between a response and an explanatory variable $(R^2_{xy} = r^2_{xy})$ represents the fraction of the total variance around the mean value $\bar{y}$ that is explained by the linear relation between x_i and y. Therefore, using (R^2_{xy}) as a variable ranking criterion enforces a ranking according to goodness of linear fit of individual variables. However, Pearson correlation measures only linear dependency between variables [22].

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is non-linear rank based non-parametric test of correlation. Its value is between -1 and 1 and interpreted in the same way as Pearson correlation coefficient for ranked variables. Spearman's rank correlation coefficient state an alternative hypothesis of the correlation between two variables corresponds to a monotonic function.

Stepwise variable selection

Backward elimination or Forward selection or Stepwise elimination can be used to select variable in the model. Backward elimination starts using all variable and variables with high P-value or above critical value are removed until the rest are significant. Forward selection starts with no variable and the variable not in the model with P-value less than critical value are inserted until the left are not significant. Stepwise elimination is the combination of them, variables are added or removed earlier in the process and the process chose the best collection of variable which maximize model fitness. Stepwise elimination is not exactly dependent on P-value rather it consider the importance of the variable in the model, this results the method to be more power full in prediction. Hence, Stepwise elimination is used to measure the interaction effect of predictors on response variable base on minimum AIC criterion [23].

Lasso variable selection

Lasso minimises the residual sum of square subject to the sum of the absolute value of the coefficients less than a constant. Lasso is help full to improve prediction accuracy by reducing large variance made by OLS trough shrinking some coefficients to zero. In this study, lasso variable selection method is applied at optimum lambda (which is in range of 1 standard deviation of minimum lambda) [24].

Dimension reduction methods

Principal factor and factor score analysis

Principal component analysis is helpful to describes the variance-covariance structure between the set of variables through a few uncorrelated new latent variables called principal components. However, the lack of correlation between principal components dose not reflect the natural correlation present on represented real variables. Therefore, a method that allow relatively slight correlation between components, like factor analysis, is preferable. factor analysis is can be applied after the number of components needed is decided to construct principal factors and factor scores. The decision for number of principal component needed can be done by considering the bend of the scree plot for principal components variances, the variance or eigenvalue of the principal component greater than one, the proportion of the total variation a counted by principal components, and subject matter consideration on principal factors composition [1]. The factor model for the random variables vector $\mathbf{Y}' = [Y_1, Y_2, \ldots Y_p]$ with mean vector μ and covariance matrix Σ is given as follow:

$$\mathbf{Y} = \mu + \Lambda\mathbf{F} + \varepsilon,$$

where Λ is $p \times k$ matrix of unknown constants called loadings, $\mathbf{F}$ is a $k \times 1$ vector of common factors and ε is a $p \times p$ diagonal matrix of specific factors. The estimates needed from this model are: covariance between factors and variables: $Cov(\mathbf{F}, \mathbf{Y}) = \mathbf{L}$ or $Cov(Y_i, F_j) = l_{ij}$, Communality: $h_i^2 = \sum_{j=1}^{k} l_{ij}^2$, and Uniqueness: $\phi_i = Var(Y_i) - h_i^2$ for $i,\ j = 1, 2, 3, ..., p$.

The i^{th} communality (h_i^2) indicates the portion of the variance of Y_i explained by k common factors and i^{th} uniqueness (ϕ_i) indicates the portion of variance of Y ($Var(Y_i)$) explained by the i^{th} specific factors. Estimated principal factors are constructed by linear combination of variables and their corresponding loadings.

$$\hat{pf}_j = \sum_{i=1}^{p} l_{ij} Y_i$$

From the result of factor model the estimated factor scores are also constructed by linear combination of original variables having relatively large loading on the factor.

$$\hat{f}_j = \sum_{i=1}^{p} l_{ij} Y_i$$

where $l_{ij} = 1$ if the variable i have relatively large loading on the factor j, else $l_{ij} = 0$ [2].

Model

Linear regression

For the data consist of a random response variable Y (number of employer in an enterprise) and k = 81 fixed explanatory variables, $X_1, X_2, \ldots, X_k$ with sample of size n = 179, linear regression is used to fit the parameter estimates and find out influential factors which determine number of employer in an enterprise. The relationship between Y and $X_1,\ X_2, \ldots, X_k$ is formulated as a linear model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are constants referred to as regression model coefficients and ϵ is a random disturbance.

It is assumed that Y is approximately a linear function of the $X's$, and ϵ measures the discrepancy in that approximation or ϵ contains no systematic information for determining Y that is not already captured by the $X's$ [25].

Logistic regression

Enterprise development status is a binary response variable with measured values Y = 1 (achieved expected progress) or Y = 0 (not achieved expected progress). Which is modelled by logistic regression model. This model is used to show the relationship between p(y) and x's for the random component have binomial distribution where $0 \leq p(y) \leq 1$.

The mean and variance of the p(y) is np and $np(1-p)$ respectively, where

$$p(y) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}$$

Logistic regression makes no assumption about the distributions of the independent variables. They do not have to be normally distributed, linearly related or of equal variance with each group. In this study, logistic regression is used to find out the influential factors from suggested predictors of enterprise development status. The influence of determinant factors are assessed individually and component wise on enterprise development status. It is modelled as follow:

$$logit(p) = \beta_0 + \beta X \tag{2}$$

where X is a matrix of independent variable, or principal factors, or factor scores in the model, β is vector of coefficients of the model, and β_0 is intercept of the model [26].

Result and discussion

Linear regression result for the number of employment using all 81 literature suggested factors showed in Appendix: Tables 12, 13, 14, 15, 16 and 17reveals that only 10 variables are significant (those are, h4, h3, IF4, IF8, Grouping, X15.29, X50.65, ed0, ed1, and *emp_male*) with 0.9992 adjusted R-squared, and similarly the result in Appendix: Tables 12, 13, 14, 15, 16 and 17 for logistic regression of development status of enterprise indicates none of the predictors are significant out of 81 factors. To address this problem variable selection and dimension reduction methods are applied to find out the real predictors of a response by removing variable redundancy, and complexity of having large number of variable.

Variable selection

The result of tests for the relation between number of employment in enterprise and predictors indicated in Table 2 reveals that, number of employers in an enterprise is significantly related at 95% confidence level with 40 explanatory variables out of 81 predictors listed in Appendix: Tables 12, 13, 14, 15, 16 and 17. Specifically, this result suggested that, as the number of employer in an enterprise increase, employment by gender is proportional, employment by eduction category is also significantly increased mainly employer with primary education is employed largely, and employment by age category

Table 2 Relation between number of employment in enterprise and explanatory variables

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
1	c11	0.465	0.607	0.417	0.397	0.432	0.487	0.677	0.755	0.816	0.094	0.012	0.874	0.103	0.169	201.759	0
2	c10	0.456	0.49	0.339	0.32	0.313	0.447	0.619	0.701	0.78	0.104	0.018	0.814	0.126	0.094	202.314	0
3	h2	0.383	0.308	0.303	0.318	0.303	0.187	0.196	0.258	0.328	0.091	− 0.042	0.574	0.076	0.31	258.068	0
4	h6	0.402	0.54	0.352	0.367	0.387	0.467	0.638	0.756	0.804	0.097	0.018	0.807	0.109	0.146	209.925	0
5	h9	0.387	0.51	0.305	0.28	0.297	0.411	0.583	0.734	0.753	0.097	0.022	0.771	0.126	0.093	208.933	0
6	h10	0.467	0.496	0.385	0.353	0.321	0.488	0.618	0.79	0.82	0.097	0.018	0.807	0.109	0.146	209.925	0
7	h12	0.423	0.493	0.32	0.304	0.296	0.44	0.597	0.747	0.776	0.097	0.022	0.771	0.126	0.093	208.933	0
8	IF8	0.322	0.085	0.038	0.022	0.011	0.002	0.007	0.008	0.009	0.094	0.186	0.012	0.245	0.001	202.591	0
9	IF10	0.693	0.314	0.243	0.151	0.108	0.019	0.008	0.012	0.007	0.086	0.193	0.01	0.161	0.032	143.065	0
10	StartK	0.272	0.295	0.324	0.346	0.352	0.271	0.282	0.335	0.426	0.123	0.061	0.42	0.117	0.119	1,154.961	0
11	CurrK	0.181	0.108	0.108	0.139	0.166	0.324	0.37	0.491	0.533	0.202	0.225	0.002	0.147	0.05	1,351.639	0
12	Category	0.215	0.076	0.043	0.025	0.033	0.032	0.094	0.268	0.36	0.17	− 0.185	0.013	− 0.205	0.006	174.274	0
13	Grouping	0	0	0	0	0	0	0	0	0	0.26	0.299	0	0.425	0	67.299	0
14	Emp_0	0.029	0.018	0.019	0.027	0.045	0.112	0.167	0.217	0.226	0.307	0.539	0	0.455	0	1,359.487	0
15	X15.29	0	0	0	0	0	0	0	0	0	0.322	0.513	0	− 0.283	0	1,024.867	0
16	X30.49	0	0	0	0	0	0	0	0	0	0.393	0.732	0	0.573	0	764.325	0
17	ed0	0.271	0.001	0	0	0	0	0	0	0	0.167	0.501	0	0.312	0	454.007	0
18	ed1	0	0	0	0	0	0	0	0	0	0.355	0.762	0	0.563	0	847.486	0
19	ed2	0.073	0.027	0.013	0.018	0.014	0.024	0.04	0.052	0.118	0.173	0.295	0	− 0.082	0.275	703.671	0
20	ed3	0	0.001	0	0	0	0	0	0	0	0.171	0.569	0	0.093	0.214	909.51	0
21	emp_Male	0.071	0	0	0	0	0	0	0	0	0.399	0.502	0	0.643	0	686.159	0
22	emp_Female	0	0	0	0	0	0	0	0	0	0.372	0.516	0	0.578	0	580.221	0
23	IF9	0.492	0.294	0.2	0.132	0.096	0.034	0.019	0.037	0.033	0.09	0.099	0.187	0.163	0.029	95.354	0.008
24	IF5	0.666	0.722	0.737	0.67	0.788	0.819	0.76	0.65	0.592	0.057	− 0.072	0.34	0.004	0.956	110.908	0.009

Table 2 (continued)

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
25	X.65	0.281	0.619	0.456	0.463	0.375	0.457	0.639	0.728	0.749	0.049	0.075	0.318	0.077	0.306	27.846	0.01
26	s1	0.64	0.571	0.504	0.556	0.569	0.603	0.362	0.211	0.122	0.086	− 0.037	0.623	− 0.101	0.177	45.297	0.011
27	IF14	0.371	0.277	0.193	0.223	0.172	0.308	0.341	0.231	0.196	0.095	0.162	0.031	0.102	0.173	91.259	0.018
28	MSEs	0.447	0.16	0.095	0.077	0.05	0.084	0.084	0.131	0.16	0.097	0.125	0.096	0.124	0.099	25.267	0.021
29	s2	0.25	0.079	0.037	0.011	0.015	0.005	0.004	0.002	0.009	0.113	− 0.091	0.226	− 0.226	0.002	23.177	0.04
30	X50.65	0.369	0.266	0.282	0.309	0.344	0.228	0.237	0.248	0.286	0.103	0.01	0.895	0.217	0.003	39.909	0.04
31	h8	0.539	0.492	0.606	0.506	0.544	0.502	0.473	0.514	0.508	0.093	0.037	0.626	0.09	0.231	158.777	0.044
32	ln4	0.34	0.108	0.102	0.081	0.069	0.019	0.009	0.007	0.01	0.087	− 0.116	0.123	− 0.242	0.001	33.375	0.152
33	h4	0.246	0.142	0.071	0.08	0.063	0.041	0.042	0.062	0.109	0.087	0.233	0.002	0.249	0.001	19.874	0.798
34	ln5	0.257	0.119	0.077	0.043	0.031	0.008	0.002	0	0.002	0.091	0.01	0.889	− 0.218	0.003	36.426	0.084
35	ln3	0.361	0.344	0.253	0.161	0.121	0.057	0.03	0.028	0.037	0.084	− 0.1	0.182	− 0.193	0.01	30.885	0.233
36	ln7	0.551	0.458	0.404	0.404	0.337	0.235	0.333	0.463	0.574	0.079	− 0.119	0.112	− 0.175	0.019	25.07	0.515
37	ln6	0.584	0.565	0.445	0.494	0.413	0.177	0.139	0.149	0.175	0.076	− 0.138	0.065	− 0.159	0.033	36.994	0.075
38	h3	0.355	0.482	0.478	0.294	0.32	0.276	0.306	0.31	0.337	0.086	0.113	0.132	0.158	0.035	22.815	0.643
39	s4	0.307	0.283	0.323	0.239	0.168	0.149	0.102	0.093	0.105	0.08	− 0.016	0.833	− 0.15	0.045	15.203	0.295
40	f5	0.307	0.222	0.165	0.152	0.166	0.247	0.377	0.584	0.678	0.098	0.16	0.032	0.137	0.068	27.46	0.386

is significantly increased for category between 30–49 and 50–65. But, the number of employer between age category 15 to 29 is decreases as the number of employer in an enterprise increases. Enterprise created by group, employer taking specific education or training on entrepreneurship, employer graduate from TVET are significantly directly correlated with the growth of enterprise's employability. Apparently, having relation with entrepreneurs for advise like as friend and any one in contact is negatively correlated with number of employment in an enterprise. The result also indicate current capital and Government investment policy motivation by Land are significantly directly correlated with number of employment in an enterprise. The influence of religion, traditionalism (cultural tackle), problems related to the legal licensing, telecommunication problems, and lack of necessary and timely marketing information have significant direct correlation with the number of employers in an enterprise. The problem of keep up with literature, get information from customers, get information from suppliers, get information from banks, and get information from commercial cooperation is higher as number of employment in an enterprise increases. The development status of an enterprise have significant have negative correlation with the number of employers in an enterprise. In addition, Starting capital, educational level, experience in self-employment, managerial experience, financial experience (financing the business), experience in the sector, firm duration, experience in business, corruption, number of employers on age category above 65, having entrepreneurs in the family, type of MSEs (micro or small), and experience as an employee have significant association with number of employment in an enterprise.

CANOVA helps to detect the relation exist between a continuous and categorical variable (only CANOVA with $k = 10$ detects type of MSEs has significant correlation with number of employment in an enterprise increases, and CANOVA have high power to detect the correlation exist between In5 (get information from suppliers) and number of employment in an enterprise increases). However, almost all significant variables detected by CANOVA are detected by Pearson or Spearman's correlation coefficient, mainly by Spearman's correlation coefficient. MIC also detects some non linear relation between some continuous variable with high power (Currk, Emp_0, $X15.29$, $X30.49$, *emp_male*, and *emp_Female*.

The result of tests for the relation between the development status of an enterprise and explanatory variables indicated in Table 3 reveals that, the development status of enterprise is significantly related at 95% confidence level with 28 explanatory variables out of 81 predictors listed in Appendix: Tables 12, 13, 14, 15, 16 and 17. This result specifically suggested that, enterprise created by group, employer with age between 15 to 29, employer taking specific education or/and training on entrepreneurship, and employer graduate from TVET are significantly directly correlated with the development of an enterprise's. The development status of an enterprise is directly significantly correlated with level of education, an enterprise with employer graduated from high school, collage or University. The influence of religion, and electric power or energy problem also increases with development status of an enterprise. The influence of availability of raw material, fear of failure, environmental conditions, problems related to the legal licensing are less on development of an enterprise. The development status of an enterprise have significant direct correlation with the current number of employers in an enterprise

Table 3 Relation between development status of an enterprise and explanatory variables

No.	X's	CANOVA P-value									MIC	Persons Correlation		Spearman Correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
1	c1	0.148	0.033	0.015	0.009	0.003	0.002	0	0	0	0.067	− 0.298	0	− 0.298	0	14.692	0
2	CurrK	0.079	0.013	0.012	0.008	0.02	0.021	0.037	0.058	0.062	0.616	0.368	0	0.682	0	125.355	0
3	MSEs	0	0	0	0	0	0	0	0	0	0.397	0.693	0	0.693	0	82.84	0
4	Category	0	0	0	0	0	0	0	0	0	0.317	− 0.567	0	− 0.621	0	70.996	0
5	Grouping	0.005	0	0	0	0	0	0	0	0	0.13	0.422	0	0.422	0	30.015	0
6	c2	0.473	0.315	0.351	0.344	0.369	0.349	0.335	0.31	0.412	0.063	0.131	0.081	0.176	0.018	15.532	0.001
7	h5	0.206	0.042	0.035	0.033	0.026	0.053	0.094	0.114	0.112	0.08	0.288	0	0.298	0	18.284	0.001
8	c15	0.38	0.3	0.28	0.226	0.244	0.203	0.193	0.179	0.252	0.064	− 0.163	0.029	− 0.165	0.028	14.345	0.003
9	Emp_0	0.231	0.251	0.284	0.216	0.275	0.295	0.408	0.382	0.469	0.13	0.091	0.226	0.36	0	28.533	0.003
10	ed2	0.15	0.12	0.122	0.118	0.14	0.176	0.219	0.27	0.276	0.097	0.241	0.001	0.214	0.004	21.402	0.003
11	IF8	0.414	0.411	0.418	0.534	0.445	0.407	0.42	0.5	0.45	0.085	0.099	0.188	0.151	0.044	17.297	0.004
12	X15.29	0.154	0.134	0.171	0.157	0.144	0.186	0.22	0.307	0.315	0.097	0.168	0.025	0.032	0.669	21.437	0.006
13	h14	0.304	0.157	0.169	0.081	0.076	0.045	0.063	0.061	0.122	0.034	0.213	0.004	0.213	0.004	7.156	0.008
14	IF2	0.271	0.145	0.09	0.087	0.069	0.041	0.051	0.062	0.091	0.069	− 0.22	0.003	− 0.245	0.001	15.464	0.009
15	f5	0.453	0.311	0.245	0.208	0.203	0.162	0.261	0.317	0.337	0.038	0.165	0.027	0.204	0.006	9.302	0.01
16	StartK	0.134	0.142	0.198	0.118	0.168	0.224	0.243	0.271	0.322	0.28	0.207	0.005	0.478	0	69.341	0.011
17	h3	0.44	0.335	0.285	0.222	0.247	0.246	0.234	0.285	0.307	0.029	0.147	0.05	0.172	0.021	6.765	0.034
18	In4	0.579	0.542	0.545	0.474	0.544	0.528	0.448	0.482	0.38	0.025	− 0.048	0.525	− 0.04	0.594	6.187	0.045
19	IF7	0.286	0.161	0.101	0.061	0.047	0.041	0.049	0.048	0.095	0.048	− 0.226	0.002	− 0.225	0.002	11.099	0.05
20	c11	0.5	0.463	0.448	0.373	0.437	0.444	0.472	0.4	0.426	0.068	0.101	0.178	0.166	0.027	15.043	0.131
21	h11	0.393	0.36	0.31	0.272	0.248	0.194	0.185	0.203	0.247	0.027	0.15	0.044	0.162	0.03	6.311	0.097
22	IF6	0.416	0.326	0.178	0.26	0.179	0.179	0.201	0.264	0.264	0.029	− 0.161	0.031	− 0.162	0.03	6.346	0.386
23	IF12	0.408	0.267	0.292	0.204	0.209	0.141	0.12	0.118	0.142	0.026	0.164	0.028	0.161	0.031	6.308	0.277
24	h6	0.416	0.457	0.462	0.451	0.525	0.416	0.466	0.462	0.463	0.066	0.091	0.225	0.159	0.034	14.521	0.151
25	h10	0.482	0.45	0.49	0.413	0.46	0.343	0.461	0.415	0.435	0.066	0.091	0.225	0.159	0.034	14.521	0.151
26	h4	0.379	0.332	0.286	0.26	0.241	0.172	0.123	0.137	0.156	0.021	0.154	0.04	0.149	0.047	4.795	0.091
27	ed3	0.182	0.124	0.132	0.13	0.148	0.209	0.242	0.241	0.257	0.066	0.214	0.004	0.129	0.084	14.237	0.076
28	h13	0.344	0.356	0.327	0.253	0.188	0.167	0.232	0.317	0.332	0.019	0.153	0.041	0.133	0.075	4.242	0.237
29	c4	0.04	0.098	0.155	0.194	0.267	0.284	0.424	0.376	0.38	0.015	− 0.128	0.087	− 0.128	0.087	0.981	0.322

and even at the start-up. The development of an micro enterprise enterprise is better than small enterprise. There is also an evidence of starting a business in group could bring a better development than an individual owned business, similarly male owned enterprises are more successful. Government investment policy motivation by land has also direct significant correlation with development of an enterprise. So government investment policy motivation is helpful for success of an enterprise. Having experience in the sector (your business), financial experience (financing the business), working by business plan, employment growth goal (the desire/want to employee), managerial skills, and experience in business have direct significant correlation with development of an enterprise. Mainly, formal managerial skills and financial experience have significant correlation with the development of an enterprise. In addition, bad experience of own have significant association with the development of an enterprise. The result indicated that, only CANOVA for $k = 2$ find out entrepreneurs activeness on business services is significantly negatively correlated with development status of an enterprise. MIC detected some non-linear relation with high power (Currk, MSEs, and Category). However, almost all significant variables detected by CANOVA are detected by Pearson or Spearman's correlation coefficient, mainly by Spearman's correlation coefficient.

Conclusion based on statistical power, the result from association and correlation analysis suggested that, CANOVA more efficiently detects continuous–continuous, and continuous-categorical non-linear or non-monotonic relation. Spearman's correlation coefficient more efficiently detects a continuous–continuous or a continuous-categorical monotonic relationship. Pearson correlation coefficient more efficiently detects the relation between continuous variables. MIC more efficiently detects non-linear or non-monotonic continuous-continuous relation. Chi-square test of independence efficiently detects relation between a continuous with a continuous, and categorical with categorical variables, but the non linear or non monotonic relation between a continuous with a categorical are not well detected. On the other hand, the results from stepwise and lasso variable selection method in Table 5 shows that, 31 variables are detected significantly as predictor for number of employment in an enterprise, and from which eleven of them are new predictors comparing to the result in association and correlation methods given in Table 2. The result using this method in Table 7 also indicates that 21 variables are significantly detected as predictors for development status of an enterprise and from which eleven of them are new predictors comparing to the result in association and correlation methods given in Table 3. Since, association and correlation can not detect the relation due to interaction effect. Similarly, some of non-causal relation between a predictor and response are not detected by lasso and stepwise variable selection methods are detected by correlation and association methods. Specifically, twenty new variables are selected as predictor for number of employment in an enterprise and nineteen new variables are selected as predictor for development status of an enterprise.

Model result from selected variables

Linear regression

1. Influencing factors affecting number of employment in an enterprise are assessed based on casual linear relation with significantly related (correlated or/union associated) predictors Table 2. Significant variables are selected based on Stepwise

elimination with minimum AIC criterion, and by lasso variable selection method. Stepwise elimination bring less number of significant variables comparing to lasso variable selection. However, both method have their own input, stepwise elimination brings three new variables (ed1, ed3, h3) those are not significant by lasso, and lasso method also brings five new variables (h2, Category, Emp_0, ed2, number of employer from 50 to 65) those are not significant by stepwise elimination. The selected variables by both methods are separately modelled, and the result in Table 4 reveals IF8, grouping, number of employer from age 15–29 and 30–49, *emp_male*, *emp_female*, and h4 are significant for both methods, where ed0, ed1, h3, and number of employer aged above 65 are only significant by stepwise elimination, similarly h2 and number of employer from age from 50 to 65 are only significant by lasso method. Finally, the variables selected by both methods are merged and the result for reduced model reveals a greater number of significant variables with equivalent model fitness as indicated in Table 4. The significance of all variables included in reduced model, unlike the lasso and stepwise selected variables, is an indication of lower multicollinearity between incorporated variables. This implies that, the predictors of number of employment in an enterprise should be the selected variable in reduced model.

2. Here, influencing factors affecting number of employment in an enterprise are assessed using all literature suggested factors in Table 5 by regression method (stepwise elimination and lasso variable selection). Significant factors are selected based on Stepwise elimination with minimum AIC criterion, and by lasso variable selection method at optimum lambda (which is in range of 1 standard deviation of minimum lambda). Unlike, the above result Table 4, regression of variables selected by stepwise elimination brings more number of significant variables comparing to variables selected by lasso method. However, both method have their own input in variable selection, stepwise elimination bring threaten new variables (c3, c6, h3, IF1, s3, s4, In1, In2, In3, In7, In10, ed1, and ed3), where five of them are not significant, but the removal of insignificant variables (IF1, s3, s4, In7 and In10) result in reduction of multiple R-squared and adjusted R-squared from 0.9946 to 0.9942, and 0.9937 to 0.9935 respectively. In addition, two significant variables In1 and In3 become insignificant. So these variables are potential variable and have to stay in the model. On the other hand, lasso method brings eight new variables (h2, h14, IF3, Category, Emp_0, $X30.49$, ed2, *emp_female*) of which three of them are only significant. The removal of insignificant variables (h14, IF3, Category, Emp_0, and ed2), resulted in reduction of multiple R-squared and adjusted R-squared from 0.9938 to 0.9932, and 0.9931 to 0.9928 respectively. However, there is no significant variable became insignificant due to the removal of those variables. This is an indication that stepwise elimination considers the gain due to interaction effect but it can result in multicollinearity, where as lasso method removes multicollinearity and the gain due to interaction effect is not considered. Due to the advantages of lasso method on controlling multicollinearity and stepwise elimination in considering interaction effect, variables selected by both stepwise elimination and lasso method are merged, and the result for reduced model reveals a greater number of significant variables with equivalent model fitness as indicated in Table 5.

Table 4 Linear regression result for number of employer in an enterprise based on selected factors through association or/union correlation methods

Variable selected by stepwise selection					Variable selected using lasso method					Reduced model					
Coefficients	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**	**Coefficients**	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**	**Coefficients**	**Estimate**	**Std. Error**	**t value**	**Pr(> |t|)**	**Significance**
Intercept	0.032	0.055	0.58	0.562	(Intercept)	0.240	0.152	1.583	0.115	(Intercept)	0.021	0.055	0.389	0.698	***
IF8	− 0.192	0.036	− 5.356	0	h2	− 0.097	0.046	− 2.109	0.037	ed1	− 0.495	0.112	− 4.406	0.000	**
IF10	− 0.042	0.028	− 1.491	0.138	IF8	− 0.187	0.036	− 5.135	0.000	h3	− 0.243	0.082	− 2.947	0.004	***
Grouping	0.421	0.088	4.761	0	IF10	− 0.046	0.029	− 1.581	0.116	IF8	− 0.201	0.034	− 5.913	0.000	***
X15.29	1.002	0.021	47.656	0	Category	− 0.011	0.031	− 0.341	0.733	Grouping	0.409	0.089	4.600	0.000	***
X30.49	1.458	0.105	13.942	0	Grouping	0.358	0.114	3.144	0.002	X15.29	0.980	0.017	56.088	0.000	***
ed0	0.399	0.129	3.097	0.002	Emp_0	0.002	0.006	0.335	0.738	X30.49	1.444	0.104	13.841	0.000	***
ed1	− 0.503	0.113	− 4.466	0	X15.29	0.968	0.025	38.188	0.000	ed0	0.386	0.130	2.975	0.003	**
ed3	− 0.053	0.033	− 1.637	0.104	X30.49	0.989	0.012	84.618	0.000	emp_Male	0.505	0.101	5.011	0.000	***
emp_Male	0.495	0.101	4.922	0	ed0	− 0.045	0.108	− 0.416	0.678	emp_Female	1.565	0.103	15.241	0.000	***
emp_Female	1.589	0.102	15.52	0	ed2	0.026	0.026	0.982	0.328	X.65	0.419	0.201	2.087	0.038	*
X.65	0.469	0.200	2.343	0.02	emp_Male	0.944	0.047	19.885	0.000	h4	0.321	0.085	3.770	0.000	***
h4	0.325	0.085	3.832	0	emp_Female	1.070	0.049	21.839	0.000						
h3	− 0.216	0.083	− 2.615	0.01	X.65	0.330	0.206	1.600	0.112						
					X50.65	0.510	0.116	4.388	0.000						
					h4	0.253	0.083	3.061	0.003						
AIC for stepwise elimination			− 267.3		Optimal lambda			0.047		Multiple R-squared			0.993		
Multiple R-squared			0.993		Multiple R-squared			0.993		Adjusted R-squared			0.993		
Adjusted R-squared			0.993		Adjusted R-squared			0.993		F-statistic			0.993		
F-statistic			1923		F-statistic			1612.000		P-value			< 2.2E−016		
P-value			< 2.2E−016		P-value			< 2.2E−016		Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 5 Linear regression for the number of employer in an enterprise based on selected factors through stepwise elimination and lasso methods

Variables selected by stepwise selection						Variables selected using lasso method						Reduced model					
Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance
(Intercept)	0.315	0.187	1.683	0.094	.	(Intercept)	0.231	0.163	1.421	0.157		(Intercept)	0.376	0.198	1.898	0.060	.
c3	− 0.298	0.150	− 1.985	0.049	*	h2	− 0.100	0.045	− 2.247	0.026	*	c3	− 0.280	0.152	− 1.843	0.067	.
c6	0.183	0.078	2.347	0.020	*	h4	0.292	0.083	3.522	0.001	***	h2	− 0.050	0.050	− 0.999	0.319	
h3	− 0.311	0.080	− 3.900	0.000	***	h14	0.136	0.085	1.593	0.113		X30.49	1.951	0.042	46.677	0.000	***
h4	0.398	0.087	4.593	0.000	***	f2	− 0.190	0.094	− 2.023	0.045	*	emp_Female	2.126	0.060	35.406	0.000	***
f2	− 0.189	0.089	− 2.111	0.036	*	IF3	− 0.033	0.028	− 1.183	0.239		c6	0.190	0.078	2.425	0.016	*
IF1	− 0.035	0.022	− 1.580	0.116		IF6	− 0.027	0.020	− 1.378	0.170		h3	− 0.272	0.089	− 3.046	0.003	***
IF6	− 0.034	0.019	− 1.813	0.072	.	IF8	− 0.189	0.035	− 5.458	0.000	***	h4	0.403	0.087	4.626	0.000	***
IF8	− 0.219	0.033	− 6.583	0.000	***	Category	− 0.011	0.031	− 0.349	0.727		f2	− 0.188	0.091	− 2.073	0.040	*
s3	− 0.056	0.040	− 1.403	0.162		Grouping	0.266	0.113	2.359	0.020	*	IF1	− 0.034	0.022	− 1.555	0.122	
s4	0.116	0.072	1.616	0.108		Emp_0	0.003	0.006	0.562	0.575		IF6	− 0.033	0.019	− 1.736	0.084	.
ln1	− 0.067	0.038	− 1.778	0.077	.	X15.29	0.964	0.025	38.848	0.000	***	IF8	− 0.219	0.033	− 6.550	0.000	***
ln2	0.128	0.071	1.806	0.073	.	X30.49	0.982	0.011	87.476	0.000	***	s3	− 0.056	0.040	− 1.400	0.163	
ln3	− 0.126	0.075	− 1.675	0.096	.	X50.65	0.486	0.113	4.305	0.000	***	s4	0.120	0.072	1.669	0.097	.
ln7	0.118	0.074	1.601	0.111		X.65	0.313	0.201	1.555	0.122		ln1	− 0.066	0.038	− 1.728	0.086	.
ln10	− 0.100	0.073	− 1.373	0.172		ed0	− 0.057	0.106	− 0.539	0.591		ln2	0.120	0.072	1.672	0.097	.
Grouping	0.384	0.090	4.276	0.000	***	ed2	0.027	0.026	1.047	0.297		ln3	− 0.112	0.077	− 1.442	0.151	
X15.29	1.007	0.020	51.029	0.000	***	emp_Male	0.950	0.047	20.421	0.000	***	ln7	0.128	0.075	1.714	0.088	***
X50.65	1.543	0.098	15.827	0.000	***	emp_Female	1.071	0.048	22.330	0.000	***	ln10	− 0.110	0.074	− 1.488	0.139	
X.65	0.428	0.190	2.252	0.026	*							Grouping	0.390	0.090	4.322	0.000	***
ed0	− 1.034	0.096	− 10.763	0.000	***							X15.29	1.006	0.020	49.555	0.000	***
ed1	1.076	0.010	103.817	0.000	***							X50.65	− 0.590	0.105	− 5.625	0.000	***
ed3	− 0.058	0.030	− 1.917	0.057	.							X.65	0.378	0.197	1.922	0.056	.
emp_Male	1.953	0.041	47.527	0.000	***							ed0	0.920	0.066	13.968	0.000	***
												ed1	− 1.051	0.046	− 23.026	0.000	***
												ed3	− 0.053	0.032	− 1.633	0.104	

Table 5 (continued)

Variables selected by stepwise selection					Variables selected using lasso method					Reduced model				
Coefficients	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)Significance**	**Coefficients**	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)Significance**	**Coefficients**	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)Significance**
AIC			− 280.540		Optimal lambda			0.052						
Multiple R-squared			0.995		Multiple R-squared			0.994		Multiple R-squared			0.995	
Adjusted R-squared			0.994		Adjusted R-squared			0.993		Adjusted R-squared			0.994	
F-statistic			1231.000		F-statistic			1415.000		F-statistic			1125.000	
P-value			< 2.2E−016		P-value			< 2.2E−016		P-value			< 2.2E−016	
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Logistic regression

1. Influencing factors affecting development status of an enterprise are assessed based on casual relation of significantly related (correlated or/union associated) predictors Table 3. Significant variables in the model are selected based on stepwise elimination with minimum AIC criterion, and lasso method at minimum lambda. Stepwise elimination does brings more variables at lower AIC than lasso method. However, both method have their own input in variable selection, stepwise elimination bring 14 new variables and of eight of them are significant variables (Grouping, IF8, StartK, IF9, IF5, s1, s2, and In4), and lasso method does bring six new variables ($X15.29$, ed2, h10, IF14, and s4). The variables selected by both methods are merged and the result for reduced model reveals a greater number of variables in the model with equivalent model fitness as indicated in Table 6, and reflects that, Grouping, IF8, IF10, CurrK, StartK, IF9, IF5, s1, MSEs, s2, In4, and f5 are significant factors on development status of an enterprise where ed1, Category, h2, h10, c11, In6, h3, and h4 are potential factors.

2. Influencing factors affecting development status of an enterprise are assessed using all literature suggested factors Table 7. Significant variables in the model are selected based on stepwise elimination at minimum AIC criterion, and by lasso variable selection method at minimum lambda. As a result stepwise elimination does brings more variables at lower AIC than lasso method. However, predictors selected by lasso method are only significant. The result for reduced model contains more variable with lower AIC, but none of the variables are significant. Hence, lasso variable selection dose in better power.

As conclusion Comparison of the results for reduced linear regressions model of variables selected by association and correlation method Table 4 with variables selected by regression method Table 5 revealed that, the earlier method does bring one new variable (*emp_male*) and the latter one does bring eight new variables (those are, IF6, $X50.65$, c3, c6, In1, In2, In3, and In7) with greater adjusted R-squared. This reveals that, based on the number of significant variables and model fitness (based on adjusted R-squared value), variables selected by lasso and stepwise elimination are taken as predictors of number of employer in an enterprise, those are listed on Table 5. Specifically, number of employer in an enterprise has significant casual relation with full self-employment, previous habitat is urban, Graduated from TVET, taken specific education/training on entrepreneurship, having other income source, environmental conditions, religion, contact with entrepreneurs in networks may be socially, visiting Bazaar, taking businesses courses, reading literatures on business, get information about business from commercial cooperation, Working MSEs in group, employers with education back ground who can not read and write, and who complete primary education, high females employment, high number of employer age between 15 to 29, 30 to 49, and above 65, and low number of employer aged between 50 to 65.

On the other hand, for categorical response variable "development status of an enterprise" the result in Tables 6 and 7 indicates that, more significant number of variables are find out by association and correlation methods, where non of variables are significant by lasso and stepwise methods with some more AIC value (with more information lost). Hence, the predictors for development status of an enterprise are variables listed in Table 6. Specifically development of an enterprise status has significant casual

Table 6 Logistic regression for development status of an enterprise based on selected factors through association or/union correlation methods

Variables selected by stepwise selection						Variables selected using lasso method						Reduced model					
Coefficients	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**	**Significance**	**Coefficients**	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**	**Significance**	**coefficients**	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**	**Significance**
(Intercept)	− 0.735	2.857	− 0.257	0.797		(Intercept)	0.925	1.316	0.703	0.482		(Intercept)	− 0.735	2.857	− 0.257	0.797	
Grouping	− 6.134	2.862	− 2.143	0.032	*	X15.29	0.091	0.283	0.322	0.747		h10	5.957	5.093	1.170	0.242	
ed1	− 3.043	2.026	− 1.502	0.133		IF10	− 0.832	0.406	− 2.051	0.040	*	Grouping	− 6.134	2.862	− 2.143	0.032	*
IF8	3.209	1.421	2.258	0.024	*	ed2	0.803	0.450	1.784	0.074	.	ed1	− 3.043	2.026	− 1.502	0.133	
IF10	− 4.749	1.782	− 2.665	0.008	**	Category	− 0.920	0.241	− 3.819	0.000	***	IF8	3.209	1.421	2.258	0.024	*
Category	− 0.857	0.541	− 1.584	0.113		CurrK	0.000	0.000	0.905	0.365		IF10	− 4.749	1.782	− 2.665	0.008	**
CurrK	0.000	0.000	3.059	0.002	**	h10	0.157	0.115	1.360	0.174		Category	− 0.857	0.541	− 1.584	0.113	
h2	1.216	0.782	1.554	0.120		IF14	− 0.370	0.230	− 1.608	0.108		CurrK	0.000	0.000	3.059	0.002	**
StartK	0.000	0.000	− 3.039	0.002	**	MSEs	5.640	1.421	3.968	0.000	***	h2	1.216	0.782	1.554	0.120	
h12	5.957	5.093	1.170	0.242		ln6	− 0.491	0.598	− 0.820	0.412		StartK	0.000	0.000	− 3.039	0.002	**
c11	− 5.971	5.108	− 1.169	0.242		s4	− 1.382	0.675	− 2.046	0.041	*	c11	− 5.971	5.108	− 1.169	0.242	
IF9	2.941	1.138	2.585	0.010	**	f5	0.995	0.582	1.709	0.088	.	IF9	2.941	1.138	2.585	0.010	**
IF5	− 3.468	1.262	− 2.747	0.006	**							IF5	− 3.468	1.262	− 2.747	0.006	**
s1	4.760	1.959	2.429	0.015	*							s1	4.760	1.959	2.429	0.015	*
MSEs	34.715	11.515	3.015	0.003	**							MSEs	34.715	11.515	3.015	0.003	**
s2	− 3.446	1.759	− 1.959	0.050	.							s2	− 3.446	1.759	− 1.959	0.050	.
ln6	1.815	1.213	1.496	0.135								ln6	1.815	1.213	1.496	0.135	
ln4	− 2.472	1.147	− 2.156	0.031	*							ln4	− 2.472	1.147	− 2.156	0.031	*
f5	1.217	0.731	1.665	0.096	.							f5	1.217	0.731	1.665	0.096	.
h3	− 3.306	2.098	− 1.576	0.115								h3	− 3.306	2.098	− 1.576	0.115	
h4	− 1.954	1.379	− 1.417	0.157								h4	− 1.954	1.379	− 1.417	0.157	
AIC				84.432		AIC				106.980		AIC				84.432	
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 7 Logistic regression for development status of an enterprise based on selected factors through stepwise elimination and lasso methods

Variables selected by stepwise selection						Variables selected using lasso method						Reduced model					
Coefficients	Estimate	Std. Error	L value	Pr(> |t|)	Significance	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance
(Intercept)	− 322.676	68577.647	− 0.005	0.996		(Intercept)	1.905	0.956	1.993	0.046	*	(Intercept)	− 954.150	32304.935	− 0.030	0.976	
c1	− 552.518	33988.886	− 0.016	0.987		c1	− 1.710	0.691	− 2.473	0.013	*	IF10	− 461.056	15,574.685	− 0.030	0.976	
c5	289.965	20,527.088	0.014	0.989		h5	0.569	0.431	1.319	0.187		c1	− 1,131.226	38,004.669	− 0.030	0.976	
h1	11.445	778.170	0.015	0.988		IF7	− 0.121	0.212	− 0.569	0.569		c5	605.125	20,639.885	0.029	0.977	
h4	− 455.053	28,358.289	− 0.016	0.987		IF10	− 0.586	0.258	− 2.275	0.023	*	h1	18.827	664.254	0.028	0.977	
h5	356.473	22,839.877	0.016	0.988		IF13	− 0.572	0.294	− 1.947	0.052	.	h4	− 655.461	22,483.411	− 0.029	0.977	
h6	− 136.468	8857.694	− 0.015	0.988		MSEs	6.475	1.542	4.199	0.000	***	h5	654.079	22,067.031	0.030	0.976	
h7	142.049	9688.991	0.015	0.988		Category	− 0.905	0.21b	− 4.217	0.000	***	h6	− 230.180	8704.315	− 0.026	0.979	
IF3	147.569	12506.807	0.012	0.991								h7	281.038	10105.656	0.028	0.978	
IF9	− 202.959	13861.819	− 0.015	0.988								IF3	237. 720	7991.684	0.030	0.976	
IF11	74.875	5115.008	0.015	0.988								IF11	224.844	7969.311	0.028	0.977	
IF13	− 365.188	21395.708	− 0.017	0.986								IF13	− 543.842	18109.693	− 0.030	0.976	
IF17	112.401	9559.584	0.012	0.991								IF17	151.610	5265.601	0.029	0.977	
s1	182.913	22360.952	0.008	0.993								s1	510.642	17258.235	0.030	0.976	
s3	− 393.450	36930.442	− 0.011	0.991								s3	− 987.537	33036.343	− 0.030	0.976	
StartK	0.001	0.102	0.014	0.989								MSES	4482.055	152684.795	0.029	0.977	
MSEs	1662.705	98798.058	0.017	0.987								Category	− 666.719	22782.652	− 0.029	0.97 7	
Category	− 364.239	21327.956	− 0.017	0.986								Grouping	− 524.788	18505.604	− 0.028	0.977	
Grouping	− 418.374	27113.658	− 0.016	0.988								X15.29	141.460	4728.631	0.030	0.976	
X15.29	75.706	4414.207	0.017	0.986													
AIC				40.000		AIC				104.460		AIC				38.000	
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

relation with working MSEs in group, religion, telecommunication problems, traditionalism (cultural tackle), current capital, corruption, entrepreneurs in the family, entrepreneurs in the friends, get information from customers, government investment policy motivation by land, and status of MSEs is being small. The development of an enterprise status is potentially related with employers with primary education, category of MSEs, year of experience in business, environmental conditions, educational level, Graduated from TVET, Specific education/training on entrepreneurship, and financial experience (financing the business).

Hence, lasso and stepwise variable selection methods are suggested for continuous response variable, and association and correlation methods are suggested for categorical response variable; or alternatively, variable selection method by combing both association, correlation, and regression method can bring a better result.

Dimension reduction

Explanatory factor analysis were applied using varimax rotation on principal components to reduce variable dimension for a purpose of avoiding complexity due to having large number of variables with out losing the needed information. Based on a result indicated in Appendix: Tables 12, 13, 14, 15, 16 and 17, 11 principal components each having a minimum of variance equal to 2, which accounts for 50.8% of the total variation in data set were taken by considering the subject matter and the bend point of a scree-plot of principal components shown in Fig. 1 too. And then, factor elements with at least 0.3 score (loading) are selected. Specifically, Factor 1 is related to Human and starting capital, Factor 2 contrasts potential input of an enterprise with influencing factors, Factor 3 contrasts an enterprise getting information from partners with an idolised enterprise, Factor 4 is related to knowledge on business mainly by training, education or courses, Factor 5 contrasts own business input with partner support, Factor 6 contrasts policy related influencing functors to Human capital, Factor 7 contrasts Entrepreneurs act for success of an enterprise with Entrepreneurs social resource, Factor 8 related to number of employer in an enterprise per categories of gender, education, and age, Factor

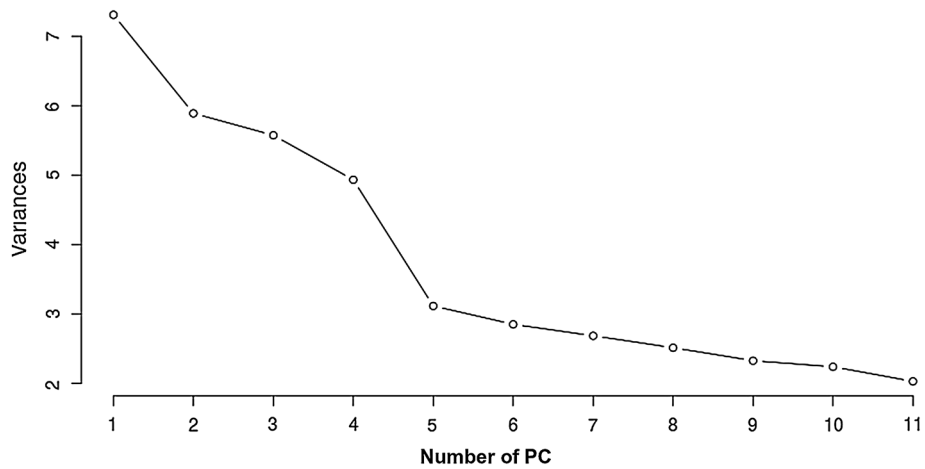

Fig. 1 Plot for number of principal components verses variance accounted

9 contrasts own contribution with partners, Factor 10 contrasts entrepreneurs nature with enterprise status, Factor 11 contrasts number of employers per category with entrepreneurs potential.

Model result for dimension reduction

Linear regression

The linear regression result for the number of employer in an enterprise based on factor scores reveals Table 8, factor 5 (contrasts of own business input with partner support), factor 6 (contrasts of policy related influencing functors to Human capital), factor 8 (variables related to Number of employer in an enterprise per categories of gender), factor 10 (contrasts of entrepreneurs nature with enterprise status), and factor 11 (contrasts number of employers per category with entrepreneurs potential) have significant affect on number of employer in an enterprise and those factors explain 82% of the variation in mean number of employer in an enterprise.

The linear regression result for number of employer in an enterprise based on principal factor reveals Table 9, principal factor 5 (contrasts of own business input with partner support), principal factor 6 (contrasts of policy related influencing functors to human capital), principal factor 7 (contrasts entrepreneurs act for success of an enterprise with entrepreneurs social resource), principal factor 9 (contrasts own contribution with partners), principal factor 10 (contrasts of entrepreneurs nature with enterprise status), and principal factor 11 (contrasts number of employers per category with entrepreneurs potential) are the significant factors those explain 85% of the variation in mean number of employer in an enterprise.

The result from regression analysis using facto score and principal factor indicates that regression analysis using principal factor gain more model fitness with one more factor. Even though, four factors are significant by both methods, factor 8 is

Table 8 Linear regression of number of employer in an enterprise based on factor scores

Full model					Reduced model by stepwise elimination				
Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Coefficients	Estimate	Std. error	t value	Pr(> |t|)
(Intercept)	0.762	1.441	0.529	0.597	(Intercept)	1.133	0.706	1.605	0.110
x1	− 0.002	0.013	− 0.115	0.909	x1	Removed			
x2	0.007	0.013	0.528	0.598	x2	Removed			
x3	0.042	0.082	0.516	0.606	x3	Removed			
x4	0.000	0.000	0.442	0.659	x4	Removed			
x5	− 0.257	0.026	− 10.063	0.000	x5	− 0.257	0.025	− 10.269	0.000
x6	0.245	0.022	11.289	0.000	x6	0.243	0.021	11.575	0.000
x7	0.014	0.096	0.144	0.886	x7	Removed			
x8	0.936	0.067	13.954	0.000	x8	0.938	0.065	14.354	0.000
x9	− 0.106	1.070	− 0.099	0.921	x9	Removed			
x10	0.000	0.000	1.864	0.064	x10	0.000	0.000	2.019	0.045
x11	0.066	0.025	2.623	0.010	x11	0.065	0.024	2.667	0.008
Multiple R-squared		0.821			Multiple R-squared		0.820		
Adjusted R-squared		0.810			Adjusted R-squared		0.815		
F-statistic		69.820			F-statistic		158.000		
P-value		< 2.2e−16			P-value		< 2.2e−16		

Table 9 Linear regression of number of employer in an enterprise based on principal factors

Full model					Reduced model by stepwise elimination				
Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Coefficients	Estimate	Std. error	t value	Pr(> |t|)
(Intercept)	0.762	1.441	0.529	0.597	(Intercept)	0.793	0.284	2.788	0.006
x1	− 0.002	0.013	− 0.115	0.909	x1	Removed			
x2	0.007	0.013	0.528	0.598	x2	Removed			
x3	0.042	0.082	0.516	0.606	x3	Removed			
x4	0.000	0.000	0.442	0.659	x4	Removed			
x5	− 0.257	0.026	− 10.063	0.000	x5	0.132	0.035	3.765	0.000
x6	0.245	0.022	11.289	0.000	x6	0.270	0.021	12.963	0.000
x7	0.014	0.096	0.144	0.886	x7	0.067	0.025	2.741	0.007
x8	0.936	0.067	13.954	0.000	x8	Removed			
x9	− 0.106	1.070	− 0.099	0.921	x9	− 0.134	0.023	− 5.781	0.000
x10	0.000	0.000	1.864	0.064	x10	1.066	0.041	25.716	0.000
x11	0.066	0.025	2.623	0.010	x11	0.313	0.056	5.612	0.000
Multiple R-squared		0.853			Multiple R-squared		0.852		
Adjusted R-squared		0.844			Adjusted R-squared		0.847		
F-statistic		88.280			F-statistic		164.700		
P-value		< 2.2e−16			P-value		< 2.2e−16		

only significant by factor score based regression, and factor 7 and 9 are only significant by principal factor based regression. Since, the result from principal factor based regression brings little gain in model fitness with complex composition (since it consider all variables than factor scores, that makes difficult to relate principal factors to real component) comparing to factor score based regression, the factor score based regression is more preferable.

Logistic regression

The logistic regression result for development status based on factor score reveals Table 10, factor 4 (related to knowledge on business mainly by training), factor 7 (contrasts Entrepreneurs act for success of an enterprise with Entrepreneurs social resource), and factor 10 (contrasts of entrepreneurs nature with enterprise status) are the significant factors with AIC of 183.16.

The logistic regression result for principal factor of development status reveals Table 11, principal factor 2 (contrasts potential input of an enterprise with Influencing factors), principal factor 3 (contrasts an enterprise getting information from partners with an idolised enterprise), principal factor 8 (Variables related to Number of employer in an enterprise per categories of gender), principal factor 9 (contrasts own contribution with partners), principal factor 10 (contrasts of entrepreneurs nature with enterprise status), and principal factor 11 (contrasts number of employers per category with entrepreneurs potential) are the significant factors with AIC of 128.348.

The result from logistic regression analysis using factor score and principal factor indicates that logistic regression analysis using principal factor brings more significant factors. Principal factor based logistic regression give 6 significant factors, where factor score based logistic regression brings 3 significant factors with lower

Table 10 Logistic regression for development status of enterprise based on factor scores

Full model					Reduced model by stepwise elimination				
Coefficients	Estimate	Std. error	z value	Pr(> |z|)	Coefficients	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 2.2437	1.6134	− 1.3907	0.1643	(Intercept)	0.215	0.796	0.270	0.787
x1	0.0154	0.0136	1.1350	0.2564	x1	Removed			
x2	− 0.0181	0.0149	− 1.2202	0.2224	x2	− 0.033	0.019	− 1.714	0.087
x3	− 0.0056	0.0895	− 0.0627	0.9500	x3	0.000	0.000	1.293	0.196
x4	0.0000	0.0000	1.7467	0.0807	x4	Removed			
x5	0.1070	0.0404	2.6505	0.0080	x5	− 0.802	0.468	− 1.714	0.086
x6	− 0.0761	0.0375	− 2.0286	0.0425	x6	0.000	0.000	4.382	0.000
x7	− 0.1203	0.1006	− 1.1960	0.2317	x7	Removed			
x8	− 0.0083	0.0723	− 0.1153	0.9082	x8	Removed			
x9	− 0.4450	1.1683	− 0.3809	0.7033	x9	Removed			
x10	0.0000	0.0000	4.9160	0.0000	x10	0.000	0.000	− 3.980	0.000
x11	0.0702	0.0447	1.5701	0.1164	x11	Removed			
AIC	192.040				AIC	183.160			

Table 11 Logistic regression for development status of enterprise based on principal factors

Full model					Reduced model by stepwise elimination				
Coefficients	Estimate	Std. error	z value	Pr(> |z|)	Coefficients	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 2.897	0.933	− 3.107	0.002	(Intercept)	− 2.365	0.440	− 5.380	0.000
x1	0.017	0.016	1.023	0.306	x1	Removed			
x2	− 0.092	0.044	− 2.099	0.036	x2	− 0.110	0.041	− 2.680	0.007
x3	− 0.069	0.070	− 0.981	0.327	x3	− 0.076	0.042	− 1.822	0.068
x4	− 0.026	0.050	− 0.514	0.608	x4	Removed			
x5	− 0.091	0.114	− 0.799	0.424	x5	Removed			
x6	− 0.161	0.119	− 1.351	0.177	x6	Removed			
x7	0.138	0.129	1.070	0.285	x7	Removed			
x8	− 0.217	0.122	− 1.783	0.075	x8	− 0.201	0.080	− 2.502	0.012
x9	0.138	0.051	2.725	0.006	x9	0.090	0.029	3.092	0.002
x10	− 0.151	0.139	− 1.087	0.277	x10	− 0.258	0.095	− 2.724	0.006
x11	0.448	0.176	2.541	0.011	x11	0.523	0.125	4.171	0.000
AIC	134.680				AIC	128.340			

AIC comparatively. Hence, principal factor based logistic regression is suggestible. Therefore, principal factor is applied in dimension reduction for a response variable is development status of an enterprise, and factor score based regression is applied in dimension reduction for a response variable is number of employers in an enterprise.

Conclusion

Regression analysis result using all literature suggested factors shows that none of the predictors for development status of an enterprise are significant, and only 10 predictors for the number of employer in an enterprise are significant out of 81 factors. As a result variable selection and dimension reduction methods are applied to assess the real predictors of a response by removing variable redundancy, and complexity of having much variable. Analysis for variable selection is done using correlation and association

methods, and regression (lasso and stepwise variable selection) methods. Related variable selection using association and correlation methods based on statistical power indicates that: CANOVA is more efficiently detects the non-linear or non-monotonic correlation between a continuous–continuous and a continuous-categorical variables. As Wang et al. [20] indicates the relation between continuous variables is well detected with more power, even if the number of significantly detected variables is smaller. Where as Spearman's correlation coefficient more efficiently detects a continuous–continuous and a continuous–categorical monotonic correlation, and Pearson correlation coefficient more efficiently detects the linear correlation between continuous variables, this result is supported by literatures [20, 22]. In addition, MIC more efficiently detects a non-linear or non-monotonic relation between continuous variables [21]. More ever, Chi-square test of independence efficiently detects relation between a continuous with a continuous, and categorical with categorical variables, but the non linear or non monotonic relation between a continuous with a categorical are not well detected. Tsai et al. [27] also suggested Chi-square in pre-processing step during data mining.

The result also reveals that, the relation between the predictor and response due to interaction effect not detected by correlation and association methods are detected by lasso and stepwise variable selection methods. Specifically, eleven new predictors for the number of employment in an enterprise, and 11 new predictors for development status of an enterprise are significantly detected by lasso and stepwise variable selection methods only. Similarly, some non-causal relation between the predictor and response are not detected by lasso and stepwise variable selection methods are also detected by correlation and association methods. Specifically, twenty new variables are significantly detected as predictor for the number of employment in an enterprise and nineteen new variables are significantly detected as predictor for development status of an enterprise by correlation and association methods only. In general, as result of Tables 2 and 5 for a continuous response variable "number of employer in an enterprise", and Tables 3 and 7 for a categorical response variable " development status of an enterprise", 51 predictors for the number of employment in an enterprise, and 40 predictors for development status of an enterprise are significantly detected. The result in literature [3, 4, 6, 7] does support the methodology applied is more general and efficient in grassing possible factors.

The result mainly indicates that, regressing the response variable "number of employer in an enterprise" based on variables selected by lasso and stepwise method does bring greater model fitness (based on adjusted R-squared value) than variables selected by association and correlation methods. Similarly, regressing the response variable "development status of an enterprise" based on variables selected by association and correlation methods does bring 12 significant variables, where none of variables are significant by lasso and stepwise elimination. Hence, lasso and stepwise variable selection methods are suggested for continuous response variable "number of employment in an enterprise", and association and correlation methods are suggested for categorical response variable "development status of an enterprise"; or alternatively filtering variables by regression, correlation and association methods and merging them for further analysis is also suggestible.

On the other hand, the result from principal factor based regression for the number of employers in an enterprise shows that, the gain in model fitness is small with complex

composition comparing to factor score based regression. But the result from logistic regression analysis for development status of an enterprise using factor score and principal factor indicates that logistic regression analysis using principal factor brings more significant number of factors with smaller information lost. Therefore, principal factor is preferred and applied in dimension reduction for a categorical response variable "development status of an enterprise", and factor score is preferred and applied in dimension reduction for a continuous response variable "number of employers in an enterprise".

The comparison of results from variable selection and dimension reduction methods indicated that, variable selection methods brings more gain in model fitness than dimension reduction methods. Hence, the suggested variable selection methods are more preferred than dimension reduction methods, and applied to find out predictors and reveals the following results.

Number of employer in an enterprise has significant casual relation with full self-employment, previous habitat is urban, Graduated from TVET, taken specific education/training on entrepreneurship, having other income source, environmental conditions, religion, contact with entrepreneurs in networks may be socially, visiting Bazaar, taking businesses courses, reading literatures on business, get information about business from commercial cooperation, Working MSEs in group, employers with education back ground who can not read and write, and who complete primary education, high females employment, high number of employer age between 15 to 29, 30 to 49 and above 65, and low number of employer aged between 50 to 65.

Development of an enterprise status has significant casual relation with working MSEs in group, religion, telecommunication problems, traditionalism (cultural tackle), current capital, corruption, entrepreneurs in the family, entrepreneurs in the friends, get information from customers, government investment policy motivation by land, and status of MSEs is being small. The development of an enterprise status is potentially related with employers with primary education, category of MSEs, year of experience in business, environmental conditions, educational level, graduated from TVET, specific education/training on entrepreneurship, and financial experience (financing the business).

In general, the suggested variable selection methods are recommended when small number of variables are studied, and the suggested dimension reduction methods are recommended for large number of variant variables (Big data case).

Future work

In this paper the measures for relation between variables are suggested based on the nature of variable. The relation due to interaction effect need more efficient method than stepwise elimination method which can consider the importance of each variable interaction effect in addition to model improvement. Due to current and recent need in Big data, a general comprehensive variable filtering and selection method should be a future work.

Abbreviations
AIC: minimum information criterion; CANOVA: continuous analysis of variance; kNN: K nearest neighbour; MIC: maximal information criterion; MINE: maximal information-based non-parametric exploration; MSEs: micro and small enterprises; MSME: micro, small, and medium enterprises; MoUDH: Ministry of Urban Development and Housing; OLS: ordinary least square; PC: principal component; SMMEs: small, medium and micro enterprises; TVET: technical vocational educational training.

Authors' contributions
This research is performed by TWH. The author read and approved the final manuscript.

Acknowledgements
The author forwards his heartfelt gratitude to anonymous reviewers for their careful reading of the manuscript and their helpful comments that improve the presentation of this work. The author also thanks Debre Markos Micro and small enterprise Authority office for MSEs data source access and respected Debre Markos enterprise and business men.

Competing interests
The author declare no competing interests.

Availability of supporting data
All support data files are available.

Consent for publication
Author proves consent of publication for this research.

Appendix

See Tables 12, 13, 14, 15, 16 and 17.

Table 12 Relation between number of employer in enterprise and explanatory variables

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
1	c1	0.513	0.599	0.719	0.717	0.724	0.634	0.624	0.627	0.627	0.063	− 0.086	0.25	− 0.078	0.297	11.579	0.562
2	c2	0.656	0.573	0.551	0.464	0.523	0.641	0.671	0.707	0.67	0.074	− 0.002	0.977	0.018	0.814	40.931	0.386
3	c3	0.4	0.302	0.382	0.552	0.538	0.824	0.973	0.983	0.988	0.041	0.039	0.602	− 0.056	0.453	7.211	0.891
4	c4	1	0.091	1	0.232	1	0.558	0.741	0.643	0.729	0.009	0.015	0.84	− 0.037	0.622	1.91	1
5	c5	0.416	0.546	0.607	0.801	0.78	0.748	0.816	0.742	0.713	0.056	− 0.048	0.525	0.009	0.907	19.278	0.825
6	c6	0.528	0.603	0.661	0.625	0.756	0.636	0.513	0.408	0.422	0.042	− 0.001	0.99	− 0.044	0.558	9.48	0.736
7	c7	0.408	0.476	0.516	0.63	0.688	0.567	0.417	0.402	0.366	0.058	− 0.025	0.736	− 0.065	0.391	11.265	0.589
8	c11	0.465	0.607	0.417	0.397	0.432	0.487	0.677	0.755	0.816	0.094	0.012	0.874	0.103	0.169	201.759	0
9	c14	0.517	0.607	0.774	0.783	0.796	0.767	0.741	0.744	0.686	0.086	− 0.017	0.825	0.001	0.99	162.571	0.126
10	c15	0.483	0.48	0.53	0.567	0.685	0.73	0.377	0.236	0.186	0.083	− 0.085	0.261	− 0.03	0.694	27.216	0.922
11	c10	0.456	0.49	0.339	0.32	0.313	0.447	0.619	0.701	0.78	0.104	0.018	0.814	0.126	0.094	202.314	0
12	h2	0.383	0.308	0.303	0.318	0.303	0.187	0.196	0.258	0.328	0.091	− 0.042	0.574	0.076	0.31	258.068	0
13	h3	0.355	0.482	0.478	0.294	0.32	0.276	0.306	0.31	0.337	0.086	0.113	0.132	0.158	0.035	22.815	0.643
14	h4	0.246	0.142	0.071	0.08	0.063	0.041	0.042	0.062	0.109	0.087	0.233	0.002	0.249	0.001	19.874	0.798
15	h5	0.425	0.48	0.445	0.599	0.524	0.717	0.702	0.71	0.718	0.051	0.001	0.985	0.119	0.113	23.259	1
16	h6	0.402	0.54	0.352	0.367	0.387	0.467	0.638	0.756	0.804	0.097	0.018	0.807	0.109	0.146	209.925	0
17	h7	0.365	0.406	0.298	0.374	0.286	0.448	0.673	0.797	0.831	0.077	− 0.034	0.652	0.054	0.474	142.3	0.777
18	h8	0.539	0.492	0.606	0.506	0.544	0.502	0.473	0.514	0.508	0.093	0.037	0.626	0.09	0.231	158.777	0.044
19	h9	0.387	0.51	0.305	0.28	0.297	0.411	0.583	0.734	0.753	0.097	0.022	0.771	0.126	0.093	208.933	0
20	h10	0.467	0.496	0.385	0.353	0.321	0.488	0.618	0.79	0.82	0.097	0.018	0.807	0.109	0.146	209.925	0
21	h11	0.493	0.661	0.697	0.795	0.792	0.738	0.633	0.602	0.527	0.053	0.053	0.485	0.083	0.268	22.201	0.986
22	h12	0.423	0.493	0.32	0.304	0.296	0.44	0.597	0.747	0.776	0.097	0.022	0.771	0.126	0.093	208.933	0
23	h13	0.469	0.518	0.34	0.324	0.288	0.174	0.2	0.2	0.19	0.069	0.02	0.789	0.018	0.807	35.7	0.621
24	h14	0.456	0.454	0.553	0.654	0.656	0.725	0.888	0.935	0.959	0.048	0.105	0.161	0.015	0.84	8.962	0.776
25	h15	0.574	0.528	0.647	0.598	0.644	0.658	0.707	0.774	0.803	0.053	0.012	0.869	0.04	0.593	12.54	0.484
26	f1	0.194	0.194	0.214	0.366	0.356	0.101	0.113	0.114	0.096	0.103	0.023	0.759	0.109	0.146	430.243	0.305
27	f2	0.567	0.655	0.634	0.619	0.599	0.73	0.773	0.833	0.833	0.039	− 0.01	0.893	− 0.009	0.9	9.049	0.769

Table 12 (continued)

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
28	f3	0.333	0.236	0.249	0.367	0.317	0.591	0.752	0.91	0.937	0.039	− 0.059	0.434	− 0.086	0.255	7.434	0.878
29	f4	0.306	0.257	0.24	0.312	0.32	0.763	0.963	0.993	0.998	0.05	− 0.066	0.382	0.053	0.482	9.556	0.73
30	f5	0.307	0.222	0.165	0.152	0.166	0.247	0.377	0.584	0.678	0.098	0.16	0.032	0.137	0.068	27.46	0.386
31	IF1	0.56	0.697	0.714	0.673	0.711	0.594	0.523	0.522	0.499	0.049	− 0.061	0.415	0.013	0.859	65.773	0.45
32	IF2	0.472	0.546	0.514	0.504	0.468	0.461	0.479	0.441	0.444	0.067	− 0.114	0.13	− 0.107	0.155	45.597	0.968
33	IF3	0.582	0.618	0.729	0.686	0.734	0.592	0.472	0.454	0.398	0.07	− 0.041	0.584	0.061	0.415	55.788	0.785
34	IF4	0.634	0.66	0.723	0.717	0.701	0.526	0.515	0.505	0.522	0.05	− 0.018	0.816	0.027	0.717	84.055	0.056
35	IF5	0.666	0.722	0.737	0.67	0.788	0.819	0.76	0.65	0.592	0.057	− 0.072	0.34	0.004	0.956	110.908	0.009
36	IF6	0.606	0.583	0.68	0.773	0.75	0.71	0.666	0.585	0.491	0.056	− 0.11	0.142	− 0.022	0.774	59.039	0.946
37	IF7	0.727	0.644	0.735	0.704	0.737	0.682	0.559	0.483	0.445	0.058	− 0.001	0.988	0.046	0.539	59.277	0.677
38	IF8	0.322	0.085	0.038	0.022	0.011	0.002	0.007	0.008	0.009	0.094	0.186	0.012	0.245	0.001	202.591	0
39	IF9	0.492	0.294	0.2	0.132	0.096	0.034	0.019	0.037	0.033	0.09	0.099	0.187	0.163	0.029	95.354	0.008
40	IF10	0.693	0.314	0.243	0.151	0.108	0.019	0.008	0.012	0.007	0.086	0.193	0.01	0.161	0.032	143.065	0
41	IF11	0.63	0.662	0.578	0.516	0.497	0.266	0.178	0.116	0.13	0.067	0.124	0.097	0.056	0.454	79.606	0.105
42	IF12	0.388	0.392	0.386	0.316	0.289	0.167	0.255	0.275	0.254	0.086	0.041	0.583	0.134	0.074	66.215	0.435
43	IF13	0.539	0.495	0.443	0.475	0.507	0.631	0.813	0.854	0.798	0.079	− 0.007	0.927	0.054	0.472	62.017	0.582
44	IF14	0.371	0.277	0.193	0.223	0.172	0.308	0.341	0.231	0.196	0.095	0.162	0.031	0.102	0.173	91.259	0.018
45	IF15	0.481	0.447	0.443	0.39	0.395	0.37	0.323	0.355	0.292	0.086	− 0.026	0.727	0.044	0.559	63.387	0.534
46	IF16	0.471	0.271	0.214	0.211	0.202	0.266	0.159	0.106	0.098	0.076	− 0.132	0.077	− 0.107	0.154	63.314	0.536
47	IF17	0.504	0.547	0.629	0.576	0.684	0.745	0.693	0.623	0.633	0.074	− 0.09	0.23	− 0.03	0.692	63.822	0.518
48	IF18	0.676	0.6	0.687	0.624	0.68	0.616	0.632	0.694	0.735	0.067	− 0.017	0.818	0.041	0.587	60.491	0.635
49	s1	0.64	0.571	0.504	0.556	0.569	0.603	0.362	0.211	0.122	0.086	− 0.037	0.623	− 0.101	0.177	45.297	0.011
50	s2	0.25	0.079	0.037	0.011	0.015	0.005	0.004	0.002	0.009	0.113	− 0.091	0.226	− 0.226	0.002	23.177	0.04
51	s3	0.385	0.518	0.549	0.417	0.488	0.687	0.911	0.464	0.456	0.05	− 0.004	0.962	− 0.09	0.229	14.726	0.962
52	s4	0.307	0.283	0.323	0.239	0.168	0.149	0.102	0.093	0.105	0.08	− 0.016	0.833	− 0.15	0.045	15.203	0.295
53	s5	0.434	0.386	0.525	0.621	0.67	0.539	0.503	0.434	0.338	0.059	0.029	0.703	0.057	0.45	11.65	0.557
54	s6	0.681	0.723	0.708	0.715	0.697	0.722	0.76	0.763	0.744	0.05	0.035	0.644	− 0.029	0.695	9.31	0.749

Table 12 (continued)

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
55	ln1	0.506	0.54	0.675	0.598	0.665	0.738	0.845	0.503	0.478	0.056	− 0.032	0.672	− 0.051	0.498	21.51	0.99
56	ln2	0.418	0.387	0.481	0.526	0.497	0.525	0.44	0.324	0.288	0.051	− 0.028	0.711	− 0.061	0.42	17.533	0.892
57	ln3	0.361	0.344	0.253	0.161	0.121	0.057	0.03	0.028	0.037	0.084	− 0.1	0.182	− 0.193	0.01	30.885	0.233
58	ln4	0.34	0.108	0.102	0.081	0.069	0.019	0.009	0.007	0.01	0.087	− 0.116	0.123	− 0.242	0.001	33.375	0.152
59	ln5	0.257	0.119	0.077	0.043	0.031	0.008	0.002	0	0.002	0.091	0.01	0.889	− 0.218	0.003	36.426	0.084
60	ln6	0.584	0.565	0.445	0.494	0.413	0.177	0.139	0.149	0.175	0.076	− 0.138	0.065	− 0.159	0.033	36.994	0.075
61	ln7	0.551	0.458	0.404	0.404	0.337	0.235	0.333	0.463	0.574	0.079	− 0.119	0.112	− 0.175	0.019	25.07	0.515
62	ln8	0.672	0.65	0.641	0.703	0.645	0.33	0.259	0.253	0.293	0.093	− 0.048	0.525	− 0.102	0.174	36.045	0.091
63	ln9	0.69	0.639	0.647	0.722	0.689	0.498	0.465	0.55	0.552	0.065	− 0.066	0.382	− 0.113	0.131	25.878	0.47
64	ln10	0.493	0.53	0.594	0.458	0.465	0.322	0.402	0.519	0.607	0.073	− 0.133	0.075	− 0.146	0.052	26.487	0.437
65	StartK	0.272	0.295	0.324	0.346	0.352	0.271	0.282	0.335	0.426	0.123	0.061	0.42	0.117	0.119	1154.961	0
66	CurrK	0.181	0.108	0.108	0.139	0.166	0.324	0.37	0.491	0.533	0.202	0.225	0.002	0.147	0.05	1351.639	0
67	SourceK	0.451	0.512	0.638	0.698	0.784	0.792	0.721	0.754	0.748	0.011	0.053	0.483	0.049	0.518	1.832	1
68	MSEs	0.447	0.16	0.095	0.077	0.05	0.084	0.084	0.131	0.16	0.097	0.125	0.096	0.124	0.099	25.267	0.021
69	Category	0.215	0.076	0.043	0.025	0.033	0.032	0.094	0.268	0.36	0.17	− 0.185	0.013	− 0.205	0.006	174.274	0
70	Grouping	0	0	0	0	0	0	0	0	0	0.26	0.299	0	0.425	0	67.299	0
71	Emp_0	0.029	0.018	0.019	0.027	0.045	0.112	0.167	0.217	0.226	0.307	0.539	0	0.455	0	1359.487	0
72	X15.29	0	0	0	0	0	0	0	0	0	0.322	0.513	0	− 0.283	0	1024.867	0
73	X30.49	0	0	0	0	0	0	0	0	0	0.393	0.732	0	0.573	0	764.325	0
74	X50.65	0.369	0.266	0.282	0.309	0.344	0.228	0.237	0.248	0.286	0.103	0.01	0.895	0.217	0.003	39.909	0.04
75	X.65	0.281	0.619	0.456	0.463	0.375	0.457	0.639	0.728	0.749	0.049	0.075	0.318	0.077	0.306	27.846	0.01
76	ed0	0.271	0.001	0	0	0	0	0	0	0	0.167	0.501	0	0.312	0	454.007	0
77	ed1	0	0	0	0	0	0	0	0	0	0.355	0.762	0	0.563	0	847.486	0
78	ed2	0.073	0.027	0.013	0.018	0.014	0.024	0.04	0.052	0.118	0.173	0.295	0	− 0.082	0.275	703.671	0
79	ed3	0	0.001	0	0	0	0	0	0	0	0.171	0.569	0	0.093	0.214	909.51	0
80	emp_Male	0.071	0	0	0	0	0	0	0	0	0.399	0.502	0	0.643	0	686.159	0
81	emp_Female	0	0	0	0	0	0	0	0	0	0.372	0.516	0	0.578	0	580.221	0

Table 13 Relation between development status of an enterprise and explanatory variables

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
1	c1	0.148	0.033	0.015	0.009	0.003	0.002	0	0	0	0.067	− 0.298	0	− 0.298	0	14.692	0
2	c2	0.473	0.315	0.351	0.344	0.369	0.349	0.335	0.31	0.412	0.063	0.131	0.081	0.176	0.018	15.532	0.001
3	c3	0.435	0.541	0.434	0.502	0.397	0.389	0.437	0.5	0.476	0.002	− 0.046	0.544	− 0.046	0.544	0.078	0.78
4	c4	0.04	0.098	0.155	0.194	0.267	0.284	0.424	0.376	0.38	0.015	− 0.128	0.087	− 0.128	0.087	0.981	0.322
5	c5	0.535	0.462	0.568	0.626	0.601	0.579	0.537	0.585	0.561	0.003	− 0.024	0.747	− 0.029	0.697	0.682	0.711
6	c6	0.6	0.495	0.534	0.467	0.554	0.545	0.454	0.483	0.55	0.001	− 0.032	0.668	− 0.032	0.668	0.066	0.797
7	c7	0.441	0.449	0.474	0.602	0.535	0.484	0.569	0.485	0.557	0.001	0.042	0.575	0.042	0.575	0.169	0.681
8	c11	0.5	0.463	0.448	0.373	0.437	0.444	0.472	0.4	0.426	0.068	0.101	0.178	0.166	0.027	15.043	0.131
9	c14	0.499	0.535	0.436	0.471	0.401	0.405	0.363	0.455	0.448	0.048	0.083	0.272	0.097	0.195	10.413	0.494
10	c15	0.38	0.3	0.28	0.226	0.244	0.203	0.193	0.179	0.252	0.064	− 0.163	0.029	− 0.165	0.028	14.345	0.003
11	c10	0.549	0.491	0.448	0.318	0.427	0.393	0.502	0.425	0.479	0.083	0.091	0.223	0.146	0.051	16.599	0.084
12	h2	0.548	0.508	0.444	0.488	0.378	0.507	0.466	0.461	0.414	0.05	0.058	0.44	0.081	0.28	10.687	0.058
13	h3	0.44	0.335	0.285	0.222	0.247	0.246	0.234	0.285	0.307	0.029	0.147	0.05	0.172	0.021	6.765	0.034
14	h4	0.379	0.332	0.286	0.26	0.241	0.172	0.123	0.137	0.156	0.021	0.154	0.04	0.149	0.047	4.795	0.091
15	h5	0.206	0.042	0.035	0.033	0.026	0.053	0.094	0.114	0.112	0.08	0.288	0	0.298	0	18.284	0.001
16	h6	0.416	0.457	0.462	0.451	0.525	0.416	0.466	0.462	0.463	0.066	0.091	0.225	0.159	0.034	14.521	0.151
17	h7	0.461	0.491	0.449	0.473	0.413	0.435	0.53	0.522	0.496	0.054	0.075	0.317	0.122	0.102	11.705	0.47
18	h8	0.428	0.377	0.335	0.315	0.328	0.264	0.333	0.284	0.279	0.043	0.128	0.087	0.107	0.155	9.634	0.473
19	h9	0.416	0.508	0.398	0.415	0.39	0.46	0.492	0.478	0.561	0.063	0.08	0.285	0.143	0.056	13.804	0.182
20	h10	0.482	0.45	0.49	0.413	0.46	0.343	0.461	0.415	0.435	0.066	0.091	0.225	0.159	0.034	14.521	0.151
21	h11	0.393	0.36	0.31	0.272	0.248	0.194	0.185	0.203	0.247	0.027	0.15	0.044	0.162	0.03	6.311	0.097
22	h12	0.444	0.443	0.409	0.401	0.444	0.473	0.498	0.473	0.489	0.063	0.08	0.285	0.143	0.056	13.804	0.182
23	h13	0.344	0.356	0.327	0.253	0.188	0.167	0.232	0.317	0.332	0.019	0.153	0.041	0.133	0.075	4.242	0.237
24	h14	0.304	0.157	0.169	0.081	0.076	0.045	0.063	0.061	0.122	0.034	0.213	0.004	0.213	0.004	7.156	0.008
25	h15	0.54	0.407	0.392	0.363	0.396	0.306	0.326	0.324	0.311	0.009	0.11	0.143	0.11	0.143	1.641	0.2
26	f1	0.28	0.232	0.219	0.267	0.219	0.177	0.236	0.308	0.338	0.165	0.109	0.145	0.002	0.974	44.384	0.072
27	f2	0.542	0.536	0.521	0.498	0.556	0.536	0.61	0.447	0.516	0	− 0.001	0.984	− 0.001	0.984	0	1

Table 13 (continued)

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
28	f3	0.359	0.39	0.323	0.402	0.371	0.317	0.378	0.35	0.389	0.01	0.121	0.106	0.121	0.106	1.622	0.203
29	f4	0.576	0.572	0.484	0.492	0.474	0.489	0.506	0.508	0.489	0.005	− 0.081	0.282	− 0.081	0.282	0.73	0.393
30	f5	0.453	0.311	0.245	0.208	0.203	0.162	0.261	0.317	0.337	0.038	0.165	0.027	0.204	0.006	9.302	0.01
31	IF1	0.452	0.42	0.44	0.416	0.374	0.401	0.479	0.395	0.467	0.025	− 0.101	0.178	− 0.124	0.099	6.052	0.301
32	IF2	0.271	0.145	0.09	0.087	0.069	0.041	0.051	0.062	0.091	0.069	− 0.22	0.003	− 0.245	0.001	15.464	0.009
33	IF3	0.485	0.47	0.409	0.357	0.426	0.338	0.441	0.424	0.455	0.036	− 0.105	0.163	− 0.12	0.11	7.892	0.162
34	IF4	0.546	0.501	0.523	0.478	0.525	0.517	0.549	0.464	0.549	0.009	− 0.047	0.529	− 0.063	0.399	1.93	0.859
35	IF5	0.58	0.549	0.588	0.511	0.528	0.585	0.49	0.537	0.524	0.022	0.035	0.642	0.048	0.521	4.788	0.571
36	IF6	0.416	0.326	0.178	0.26	0.179	0.179	0.201	0.264	0.264	0.029	− 0.161	0.031	− 0.162	0.03	6.346	0.386
37	IF7	0.286	0.161	0.101	0.061	0.047	0.041	0.049	0.048	0.095	0.048	− 0.226	0.002	− 0.225	0.002	11.099	0.05
38	IF8	0.414	0.411	0.418	0.534	0.445	0.407	0.42	0.5	0.45	0.085	0.099	0.188	0.151	0.044	17.297	0.004
39	IF9	0.501	0.566	0.535	0.52	0.473	0.515	0.486	0.562	0.57	0.058	0.002	0.975	0.036	0.634	10.997	0.051
40	IF10	0.406	0.419	0.498	0.403	0.413	0.47	0.452	0.492	0.581	0.052	− 0.106	0.158	− 0.044	0.557	10.242	0.069
41	IF11	0.46	0.529	0.472	0.569	0.483	0.533	0.469	0.476	0.543	0.01	− 0.06	0.424	− 0.057	0.445	2.392	0.793
42	IF12	0.408	0.267	0.292	0.204	0.209	0.141	0.12	0.118	0.142	0.026	0.164	0.028	0.161	0.031	6.308	0.277
43	IF13	0.439	0.285	0.32	0.239	0.275	0.187	0.306	0.243	0.408	0.017	− 0.146	0.052	− 0.134	0.073	3.954	0.556
44	IF14	0.433	0.464	0.37	0.347	0.329	0.301	0.341	0.425	0.44	0.025	− 0.117	0.12	− 0.096	0.202	5.675	0.339
45	IF15	0.409	0.42	0.37	0.331	0.25	0.261	0.273	0.368	0.398	0.02	− 0.142	0.057	− 0.131	0.079	4.57	0.471
46	IF16	0.572	0.506	0.571	0.518	0.526	0.529	0.487	0.546	0.559	0.022	0.045	0.547	0.035	0.643	5.348	0.375
47	IF17	0.446	0.373	0.342	0.301	0.261	0.228	0.183	0.156	0.144	0.036	0.136	0.07	0.145	0.053	8.549	0.129
48	IF18	0.495	0.616	0.555	0.505	0.593	0.565	0.497	0.579	0.537	0.008	− 0.039	0.608	− 0.031	0.68	1.942	0.857
49	s1	0.524	0.429	0.36	0.33	0.395	0.295	0.424	0.434	0.405	0.015	0.085	0.259	0.035	0.644	2.947	0.229
50	s2	0.44	0.413	0.349	0.313	0.352	0.259	0.248	0.216	0.217	0.01	− 0.12	0.11	− 0.12	0.11	2.102	0.147
51	s3	0.369	0.258	0.331	0.347	0.119	0.212	0.293	0.376	0.458	0.01	− 0.098	0.194	− 0.099	0.187	2.116	0.347
52	s4	0.445	0.408	0.371	0.361	0.297	0.284	0.247	0.263	0.243	0.01	− 0.119	0.111	− 0.119	0.111	2.088	0.148
53	s5	0.545	0.489	0.621	0.547	0.556	0.545	0.65	0.597	0.605	0	− 0.024	0.754	− 0.024	0.754	0.027	0.871
54	s6	0.462	0.464	0.445	0.47	0.42	0.421	0.409	0.381	0.32	0.005	− 0.08	0.287	− 0.08	0.287	0.844	0.358

Table 13 (continued)

No.	X's	CANOVA P-value									MIC	Persons correlation		Spearman correlation		Chi-square	
		K = 2	K = 4	K = 6	K = 8	K = 10	K = 20	K = 30	K = 40	K = 50		Value	P-value	Value	P-value	Value	P-value
55	In1	0.464	0.474	0.433	0.432	0.33	0.218	0.262	0.343	0.446	0.007	− 0.028	0.707	0.024	0.751	1.489	0.685
56	In2	0.491	0.579	0.425	0.574	0.554	0.619	0.561	0.54	0.555	0.002	0.039	0.604	0.041	0.583	0.366	0.833
57	In3	0.489	0.513	0.54	0.432	0.394	0.415	0.388	0.391	0.39	0.016	− 0.086	0.254	− 0.092	0.221	4.12	0.127
58	In4	0.579	0.542	0.545	0.474	0.544	0.528	0.448	0.482	0.38	0.025	− 0.048	0.525	− 0.04	0.594	6.187	0.045
59	In5	0.546	0.535	0.445	0.555	0.583	0.532	0.477	0.45	0.502	0.016	− 0.042	0.58	− 0.045	0.55	3.857	0.145
60	In6	0.454	0.379	0.431	0.361	0.377	0.257	0.317	0.341	0.315	0.012	− 0.124	0.098	− 0.128	0.088	3.055	0.217
61	In7	0.566	0.649	0.508	0.521	0.575	0.633	0.607	0.638	0.638	0.008	0.023	0.759	0.032	0.673	1.973	0.373
62	In8	0.544	0.591	0.574	0.587	0.56	0.542	0.531	0.525	0.474	0.003	− 0.007	0.929	− 0.012	0.87	0.82	0.664
63	In9	0.574	0.473	0.553	0.511	0.509	0.493	0.618	0.601	0.505	0.001	− 0.029	0.697	− 0.031	0.677	0.293	0.864
64	In10	0.484	0.473	0.483	0.555	0.566	0.494	0.475	0.447	0.551	0.004	− 0.054	0.471	− 0.058	0.443	0.907	0.636
65	StartK	0.134	0.142	0.198	0.118	0.168	0.224	0.243	0.271	0.322	0.28	0.207	0.005	0.478	0	69.341	0.011
66	CurrK	0.079	0.013	0.012	0.008	0.02	0.021	0.037	0.058	0.062	0.616	0.368	0	0.682	0	125.355	0
67	SourceK	0.445	0.297	0.36	0.416	0.474	0.355	0.373	0.322	0.425	0.006	− 0.095	0.205	− 0.095	0.205	0.887	0.346
68	MSEs	0	0	0	0	0	0	0	0	0	0.397	0.693	0	0.693	0	82.84	0
69	Category	0	0	0	0	0	0	0	0	0	0.317	− 0.567	0	− 0.621	0	70.996	0
70	Grouping	0.005	0	0	0	0	0	0	0	0	0.13	0.422	0	0.422	0	30.015	0
71	Emp_0	0.231	0.251	0.284	0.216	0.275	0.295	0.408	0.382	0.469	0.13	0.091	0.226	0.36	0	28.533	0.003
72	X15.29	0.154	0.134	0.171	0.157	0.144	0.186	0.22	0.307	0.315	0.097	0.168	0.025	0.032	0.669	21.437	0.006
73	X30.49	0.341	0.067	0.124	0.129	0.083	0.167	0.24	0.275	0.306	0.056	0.136	0.069	0.133	0.077	12.452	0.053
74	X50.65	0.635	0.506	0.48	0.488	0.552	0.557	0.446	0.505	0.575	0.005	− 0.057	0.449	− 0.048	0.52	0.917	0.632
75	X.65	0.284	0.605	0.463	0.484	0.454	0.536	0.444	0.501	0.561	0	0.009	0.91	0.009	0.91	0	1
76	ed0	0.445	0.41	0.44	0.379	0.371	0.325	0.339	0.343	0.387	0.043	0.018	0.815	− 0.131	0.08	8.923	0.063
77	ed1	0.332	0.281	0.223	0.087	0.094	0.204	0.273	0.309	0.361	0.054	0.113	0.13	0	0.995	11.331	0.079
78	ed2	0.15	0.12	0.122	0.118	0.14	0.176	0.219	0.27	0.276	0.097	0.241	0.001	0.214	0.004	21.402	0.003
79	ed3	0.182	0.124	0.132	0.13	0.148	0.209	0.242	0.241	0.257	0.066	0.214	0.004	0.129	0.084	14.237	0.076
80	emp_Male	0.375	0.19	0.268	0.171	0.212	0.172	0.26	0.289	0.323	0.026	0.121	0.107	0.078	0.298	5.86	0.439
81	emp_Female	0.456	0.438	0.376	0.419	0.396	0.298	0.295	0.29	0.389	0.04	0.069	0.358	− 0.049	0.518	9.036	0.108

Table 14 Principal factors

No.	X's	MR1	MR2	MR3	MR4	MR7	MR6	MR11	MR5	MR10	MR8	MR9
1	h14	0.046	− 0.409	0.118	0.191	− 0.255	0.019	0.067	− 0.091	0.03	0.035	0.16
2	h5	0.126	− 0.2	− 0.023	0.255	− 0.255	− 0.043	− 0.04	0.046	0.144	− 0.002	0.224
3	c14	0.07	− 0.197	− 0.114	− 0.181	− 0.093	− 0.06	0.127	0.078	0.065	− 0.03	0.133
4	c3	0.042	− 0.175	− 0.019	0.052	− 0.269	0.065	0.007	0.145	− 0.17	0.02	− 0.05
5	ln6	− 0.133	− 0.146	0.578	0.067	0.166	− 0.067	0.006	− 0.069	− 0.055	− 0.043	0.132
6	ln7	0.002	− 0.141	0.652	0.091	0.174	− 0.078	− 0.038	− 0.069	0.079	0.013	0.154
7	h11	0.033	− 0.134	0.132	0.759	− 0.021	0.018	0.084	0.009	− 0.074	− 0.06	0.027
8	h13	0.02	− 0.128	0.104	0.648	0.115	− 0.014	0.13	0.061	0.089	− 0.104	0.036
9	h15	− 0.016	− 0.099	0.145	0.201	− 0.056	0.078	− 0.309	0.1	0.126	0.134	0.009
10	f5	0.02	− 0.095	0.03	0.227	− 0.091	0.14	0.076	− 0.072	− 0.005	0.06	0.085
11	c15	0.015	− 0.079	− 0.127	− 0.138	− 0.223	− 0.058	− 0.091	0.007	− 0.036	− 0.017	0.085
12	MSEs	− 0.051	− 0.077	− 0.042	0.159	− 0.058	0.119	0.063	− 0.042	0.572	− 0.149	0.001
13	ln5	− 0.068	− 0.069	0.324	0.075	0.35	0.171	− 0.132	0.047	− 0.033	− 0.054	0.248
14	CurrK	− 0.044	− 0.068	0.18	0.029	0.114	0.237	0.158	− 0.038	0.451	− 0.004	− 0.082
15	c11	0.976	− 0.066	− 0.053	0.006	− 0.019	0.001	− 0.02	0.03	0.01	0	− 0.018
16	Emp_0	− 0.028	− 0.057	− 0.161	− 0.044	− 0.072	0.497	0.325	0.094	0.073	0.069	− 0.013
17	c10	0.992	− 0.05	− 0.039	0.024	− 0.017	0.012	− 0.027	0.021	0.005	0.007	− 0.027
18	h9	0.992	− 0.048	− 0.032	0.031	0.006	− 0.005	− 0.019	0.009	− 0.007	0.015	− 0.018
19	h6	0.997	− 0.04	− 0.029	0.026	− 0.006	− 0.002	− 0.03	0.013	0	0.01	− 0.021
20	h10	0.997	− 0.04	− 0.029	0.026	− 0.006	− 0.002	− 0.03	0.013	0	0.01	− 0.021
21	f2	0.022	− 0.039	− 0.027	0.051	0.385	− 0.085	0.153	− 0.126	0.056	− 0.024	− 0.007
22	Grouping	− 0.013	− 0.038	0.006	0.315	− 0.252	0.172	0.294	− 0.054	0.549	− 0.041	0.048
23	h7	0.883	− 0.036	0.006	− 0.025	0.027	− 0.017	− 0.031	− 0.037	0.018	− 0.011	0.126
24	h12	0.991	− 0.034	− 0.031	0.002	− 0.012	0.002	− 0.031	0.023	− 0.015	0.019	− 0.022
25	s4	− 0.02	− 0.033	0.048	0.063	0.501	0.075	− 0.196	0.113	0.009	− 0.005	0.041
26	ln2	0.016	− 0.033	0.318	0.305	0.166	− 0.063	0.097	− 0.148	− 0.038	− 0.023	0.147
27	emp_Female	− 0.029	− 0.032	− 0.046	0.079	0.003	0.075	0.403	0.042	− 0.02	0.485	− 0.085
28	c7	− 0.059	− 0.027	− 0.006	− 0.093	− 0.036	0.034	− 0.09	− 0.929	0.008	− 0.007	0.055

Table 14 (continued)

No.	X's	MR1	MR2	MR3	MR4	MR7	MR6	MR11	MR5	MR10	MR8	MR9
29	ed3	− 0.223	− 0.023	− 0.083	0.167	0.027	0.266	0.608	0.136	0.024	0.307	0.008
30	Category	− 0.043	− 0.018	− 0.053	− 0.374	0.101	− 0.03	− 0.145	− 0.004	− 0.574	− 0.057	− 0.203
31	s3	− 0.026	− 0.015	− 0.029	− 0.079	0.202	0.024	− 0.107	0.144	− 0.003	0.041	0.034
32	ed1	0.012	− 0.005	− 0.113	− 0.032	0.03	0.882	0.081	− 0.028	0.051	0.081	0.01
33	In8	− 0.008	− 0.001	0.746	0.001	− 0.025	− 0.037	0.021	0.078	0.018	0.028	0.025
34	X50.65	− 0.114	0.001	− 0.136	− 0.019	0.011	− 0.094	0.011	− 0.007	− 0.003	0.104	− 0.002
35	X.65	− 0.078	0.003	0.046	0.002	0.013	0.012	− 0.01	0.145	− 0.077	0.291	0.221
36	f1	0.015	0.006	− 0.165	0.322	− 0.03	− 0.034	− 0.076	− 0.099	0.043	0.037	− 0.05
37	SourceK	− 0.011	0.007	− 0.068	0.014	− 0.005	0.013	− 0.003	0.073	0.027	0.014	− 0.728
38	StartK	− 0.078	0.008	0.188	− 0.1	0.205	0.066	0.014	− 0.009	0.411	0.019	− 0.036
39	X15.29	− 0.122	0.016	0.022	0.1	− 0.06	0.082	0.837	0.148	0.017	0.142	− 0.022
40	f3	− 0.002	0.019	0.153	0.034	0.033	0.002	− 0.037	− 0.018	− 0.019	− 0.013	0.771
41	h8	0.136	0.02	− 0.074	0.06	− 0.08	0.147	− 0.044	0.126	0.139	− 0.145	0.035
42	ed2	0.008	0.022	0.048	0.124	− 0.078	0.014	0.517	− 0.006	0.177	− 0.085	− 0.032
43	c6	− 0.011	0.022	− 0.083	0.194	− 0.046	0.064	0.058	0.017	− 0.263	− 0.142	0.08
44	s5	0.062	0.024	0.008	0.099	0.011	− 0.035	0.094	0.904	− 0.002	0.011	− 0.048
45	c4	− 0.008	0.025	0.122	0.049	− 0.199	0.107	− 0.007	0.05	− 0.254	− 0.004	0.01
46	In10	− 0.055	0.034	0.84	− 0.03	− 0.017	− 0.086	− 0.017	0.072	0.025	− 0.021	− 0.066
47	ed0	0.009	0.036	− 0.029	0.03	− 0.017	0.057	0.063	− 0.007	− 0.043	0.937	0.017
48	In1	0.023	0.038	0.233	0.123	0.113	0.015	− 0.047	− 0.075	0.064	0.002	0.013
49	In9	− 0.024	0.038	0.895	− 0.014	− 0.068	− 0.079	− 0.004	0.037	0.049	0.077	− 0.023
50	s1	0.083	0.041	0.097	0.19	0.482	− 0.043	0.031	− 0.048	0.092	− 0.016	0.004
51	X30.49	0.065	0.041	− 0.116	− 0.03	0.011	0.9	− 0.103	− 0.082	0.128	0.101	0.007
52	h2	− 0.04	0.047	0.065	0.62	0.103	− 0.123	0.046	0.152	− 0.013	− 0.017	− 0.118
53	In3	− 0.072	0.058	0.387	0.246	0.314	− 0.045	− 0.041	0.013	− 0.187	− 0.051	0.109
54	In4	− 0.141	0.075	0.347	0.092	0.424	0.051	− 0.206	0.135	− 0.07	− 0.043	0.217
55	s2	0.051	0.082	0.151	0.078	0.64	− 0.112	− 0.027	0.06	− 0.024	0.034	− 0.102
56	emp_Male	0.135	0.095	− 0.013	0.073	− 0.044	0.091	− 0.06	− 0.085	0.128	0.734	− 0.058

Table 14 (continued)

No.	X's	MR1	MR2	MR3	MR4	MR7	MR6	MR11	MR5	MR10	MR8	MR9
57	s6	− 0.051	0.096	− 0.19	0.042	− 0.153	0.043	0.108	0.202	− 0.047	− 0.118	0.088
58	c1	− 0.032	0.096	0.045	− 0.117	0.213	0.025	− 0.091	0.13	− 0.208	0.006	0.048
59	h3	0.018	0.098	0.046	0.689	0.124	− 0.04	0.052	0.103	0.17	0.138	− 0.073
60	c2	− 0.02	0.105	0.111	0.289	0.166	− 0.054	− 0.16	0.202	0.208	0.138	0.176
61	IF12	0.158	0.115	− 0.104	0.089	0.056	− 0.067	− 0.159	0.16	0.424	0.064	− 0.01
62	h4	0.058	0.136	0.096	0.601	0.122	0.13	0.016	0.144	0.158	0.134	0.121
63	c5	− 0.078	0.16	− 0.051	0.209	0.178	− 0.013	− 0.129	0.157	− 0.054	− 0.02	− 0.024
64	IF16	0.122	0.188	− 0.138	− 0.025	0.169	− 0.275	− 0.098	− 0.01	0.096	0.067	0.165
65	f4	0.01	0.19	0.078	0.144	0.162	0	− 0.212	0.245	− 0.108	0.048	− 0.01
66	IF17	0.049	0.345	− 0.222	− 0.022	0.127	− 0.239	− 0.097	0.036	0.241	0.091	0.193
67	IF8	0.053	0.416	− 0.098	0.073	− 0.25	0.003	0.261	− 0.099	0.175	0.026	0.135
68	IF5	0.009	0.425	0.006	− 0.03	− 0.191	− 0.073	− 0.097	0.046	0.18	0.005	0.272
69	IF18	− 0.045	0.45	− 0.244	− 0.001	0.109	− 0.169	− 0.024	− 0.017	0.039	0.144	0.169
70	IF9	0.04	0.461	0.07	0.039	− 0.15	− 0.016	0.186	− 0.053	0.179	− 0.009	0.219
71	IF6	− 0.048	0.479	0.167	0.002	0.029	− 0.006	− 0.207	0.169	0.013	− 0.032	− 0.06
72	IF10	− 0.097	0.482	0.015	0.074	− 0.184	0.165	0.095	− 0.005	0.117	− 0.03	0.002
73	IF1	0.039	0.51	0.132	− 0.11	0.111	− 0.028	0.006	− 0.058	− 0.169	− 0.008	0.047
74	IF11	− 0.054	0.539	− 0.098	0.039	− 0.059	0.115	0.046	0.188	0.035	− 0.081	0.06
75	IF7	− 0.065	0.548	− 0.053	− 0.117	0.065	− 0.04	0.05	− 0.045	− 0.051	− 0.045	− 0.032
76	IF13	0.034	0.552	− 0.02	0.052	0.054	0.002	− 0.01	0.081	0.005	0.008	− 0.098
77	IF14	− 0.014	0.605	− 0.092	0.082	0.025	0.067	0.037	0.11	0.035	0.107	− 0.131
78	IF15	− 0.016	0.607	− 0.067	0.019	0.065	− 0.104	− 0.005	0.086	0.111	0.06	0.011
79	IF2	− 0.022	0.61	0.079	− 0.048	0.115	− 0.033	− 0.078	0.017	− 0.267	− 0.04	− 0.044
80	IF3	− 0.064	0.688	− 0.017	0.027	0.098	− 0.009	− 0.018	− 0.182	− 0.176	0.018	− 0.045
81	IF4	− 0.014	0.695	0.039	− 0.001	− 0.046	− 0.031	− 0.005	− 0.098	− 0.085	0.053	0.051

Table 15 Variance of principal components

Principal components (Pc)	Standard deviation	Proportion of variance	Cumulative proportion	Variance of Pc
PC1	2.702	0.090	0.090	7.302
PC2	2.427	0.073	0.163	5.889
PC3	2.360	0.069	0.232	5.570
PC4	2.207	0.060	0.292	4.870
PC5	1.759	0.038	0.330	3.093
PC6	1.681	0.035	0.365	2.824
PC7	1.622	0.032	0.397	2.632
PC8	1.546	0.030	0.427	2.389
PC9	1.523	0.029	0.455	2.320
PC10	1.494	0.028	0.483	2.231
PC11	1.423	0.025	0.508	2.025
PC12	1.374	0.023	0.531	1.887
PC13	1.341	0.022	0.553	1.798
PC14	1.286	0.020	0.574	1.653
PC15	1.271	0.020	0.594	1.616
PC16	1.239	0.019	0.613	1.534
PC17	1.191	0.018	0.630	1.417
PC18	1.174	0.017	0.647	1.379
PC19	1.165	0.017	0.664	1.357
PC20	1.125	0.016	0.680	1.266
PC21	1.100	0.015	0.695	1.211
PC22	1.058	0.014	0.708	1.119
PC23	1.042	0.013	0.722	1.086
PC24	1.018	0.013	0.735	1.036
PC25	1.017	0.013	0.747	1.034
PC26	0.997	0.012	0.760	0.993
PC27	0.972	0.012	0.771	0.945
PC28	0.955	0.011	0.783	0.912
PC29	0.933	0.011	0.793	0.871
PC30	0.918	0.010	0.804	0.843
PC31	0.888	0.010	0.813	0.789
PC32	0.879	0.010	0.823	0.772
PC33	0.869	0.009	0.832	0.756
PC34	0.847	0.009	0.841	0.717
PC35	0.835	0.009	0.850	0.698
PC36	0.829	0.008	0.858	0.688
PC37	0.810	0.008	0.866	0.656
PC38	0.789	0.008	0.874	0.622
PC39	0.769	0.007	0.881	0.592
PC40	0.755	0.007	0.888	0.571
PC41	0.734	0.007	0.895	0.539
PC42	0.716	0.006	0.901	0.512
PC43	0.704	0.006	0.908	0.496
PC44	0.690	0.006	0.913	0.476
PC45	0.686	0.006	0.919	0.471
PC46	0.664	0.005	0.925	0.441
PC47	0.660	0.005	0.930	0.435
PC48	0.634	0.005	0.935	0.402
PC49	0.617	0.005	0.940	0.380

Table 15 (continued)

Principal components (Pc)	Standard deviation	Proportion of variance	Cumulative proportion	Variance of Pc
PC50	0.602	0.004	0.944	0.362
PC51	0.598	0.004	0.949	0.358
PC52	0.578	0.004	0.953	0.334
PC53	0.557	0.004	0.957	0.310
PC54	0.547	0.004	0.960	0.299
PC55	0.541	0.004	0.964	0.293
PC56	0.521	0.003	0.967	0.272
PC57	0.503	0.003	0.970	0.253
PC58	0.483	0.003	0.973	0.234
PC59	0.466	0.003	0.976	0.217
PC60	0.450	0.003	0.978	0.202
PC61	0.440	0.002	0.981	0.194
PC62	0.420	0.002	0.983	0.177
PC63	0.414	0.002	0.985	0.171
PC64	0.410	0.002	0.987	0.168
PC65	0.394	0.002	0.989	0.155
PC66	0.385	0.002	0.991	0.148
PC67	0.372	0.002	0.993	0.138
PC68	0.353	0.002	0.994	0.125
PC69	0.329	0.001	0.995	0.108
PC70	0.299	0.001	0.997	0.089
PC71	0.280	0.001	0.998	0.078
PC72	0.271	0.001	0.998	0.074
PC73	0.226	0.001	0.999	0.051
PC74	0.162	0.000	0.999	0.026
PC75	0.149	0.000	1.000	0.022
PC76	0.122	0.000	1.000	0.015
PC77	0.077	0.000	1.000	0.006
PC78	0.068	0.000	1.000	0.005
PC79	0.043	0.000	1.000	0.002
PC80	0.000	0.000	1.000	0.000
PC81	0.000	0.000	1.000	0.000

Table 16 Linear regression full model for number of employer in enterprise based on all explanatory variables

No.	Full model												
	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance	No.	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance
1	(Intercept)	− 0.099	1.196	− 0.083	0.934		42	IF13	0.022	0.046	0.494	0.622	
2	c1	0.059	0.123	0.482	0.631		43	IF14	0.048	0.054	0.883	0.380	
3	c2	0.025	0.057	0.427	0.670		44	IF15	− 0.021	0.039	− 0.525	0.601	
4	c3	− 0.306	0.213	− 1.434	0.155		45	IF16	0.010	0.029	0.329	0.743	
5	c4	− 0.096	0.490	− 0.196	0.845		46	IF17	− 0.018	0.033	− 0.529	0.598	
6	c5	0.005	0.105	0.048	0.962		47	IF18	0.013	0.035	0.355	0.723	
7	c6	0.136	0.120	1.135	0.259		48	s1	0.002	0.096	0.024	0.981	
8	c7	0.309	0.446	0.694	0.490		49	s2	− 0.006	0.125	− 0.046	0.963	
9	c11	− 0.011	0.118	− 0.093	0.926		50	s3	− 0.084	0.056	− 1.505	0.136	
10	c14	− 0.017	0.028	− 0.615	0.540		51	s4	0.077	0.119	0.650	0.517	
11	c15	0.037	0.079	0.469	0.640		52	s5	0.330	0.433	0.764	0.447	
12	c10	− 0.329	0.262	− 1.255	0.212		53	s6	− 0.079	0.104	− 0.755	0.452	
13	h1	− 0.005	0.006	− 0.792	0.430		54	In1	− 0.035	0.057	− 0.615	0.540	
14	h2	− 0.092	0.076	− 1.211	0.229		55	In2	0.150	0.105	1.431	0.156	
15	h3	− 0.334	0.142	− 2.351	0.021	*	56	In3	− 0.110	0.115	− 0.959	0.340	
16	h4	0.354	0.131	2.704	0.008	**	57	In4	0.033	0.122	0.272	0.786	
17	h5	− 0.037	0.083	− 0.446	0.657		58	In5	0.035	0.110	0.317	0.752	
18	h6	0.119	0.493	0.242	0.809		59	In6	− 0.145	0.124	− 1.164	0.247	
19	h7	− 0.053	0.044	− 1.187	0.238		60	In7	0.149	0.143	1.041	0.300	
20	h8	0.004	0.020	0.216	0.829		61	In8	− 0.021	0.135	− 0.158	0.875	
21	h9	0.213	0.242	0.880	0.381		62	In9	0.170	0.335	0.508	0.613	
22	h11	0.122	0.126	0.971	0.334		63	In10	− 0.228	0.294	− 0.776	0.439	
23	h12	0.059	0.306	0.193	0.848		64	StartK	0.000	0.000	0.600	0.550	
24	h13	0.016	0.155	0.104	0.917		65	CurrK	0.000	0.000	− 0.609	0.544	
25	h14	0.140	0.137	1.024	0.308		66	SourceK	− 0.303	0.331	− 0.914	0.363	

Table 16 (continued)

No.	Full model												
	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance	No.	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	Significance
26	h15	0.023	0.141	0.163	0.871		67	MSEs	− 0.068	0.142	− 0.480	0.632	
27	f1	0.000	0.000	− 0.441	0.660		68	Category	− 0.028	0.051	− 0.561	0.576	
28	f2	− 0.210	0.136	− 1.546	0.125		69	Grouping	0.323	0.168	1.925	0.057	.
29	f3	− 0.254	0.365	− 0.696	0.488		70	Emp_0	0.011	0.010	1.122	0.265	
30	f4	− 0.060	0.153	− 0.391	0.697		71	X15.29	0.981	0.041	24.100	0.000	***
31	f5	− 0.011	0.123	− 0.090	0.929		72	X30.49	0.006	0.064	0.091	0.928	
32	IF1	− 0.025	0.045	− 0.546	0.587		73	X50.65	1.589	0.169	9.413	0.000	***
33	IF2	0.003	0.042	0.069	0.945		74	X.65	0.328	0.274	1.198	0.234	
34	IF3	− 0.042	0.051	− 0.817	0.416		75	ed0	− 1.034	0.154	− 6.732	0.000	***
35	IF4	0.034	0.052	0.662	0.510		76	ed1	1.050	0.069	15.313	0.000	***
36	IF5	0.011	0.029	0.374	0.710		77	ed2	0.011	0.035	0.314	0.754	
37	IF6	− 0.066	0.033	− 2.001	0.048	*	78	ed3	− 0.023	0.055	− 0.415	0.679	
38	IF7	0.008	0.036	0.232	0.817		79	emp_Male	1.923	0.079	24.259	0.000	***
39	IF8	− 0.245	0.052	− 4.691	0.000	***	80	IF10	− 0.020	0.049	− 0.418	0.677	
40	IF9	− 0.015	0.049	− 0.302	0.764		81	IF11	0.027	0.040	0.676	0.501	
41	IF12	0.016	0.026	0.641	0.523								
Multiple R-squared			0.995										
Adjusted R-squared			0.992										
F-statistic			261.900										
P-value			< 2.2E−016										
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1													

Table 17 Logistic regression full model for development status of an enterprise based on all explanatory variables

No.	Full model										
	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	No	Coefficients	Estimate	Std. error	t value	Pr(> |t|)
1	(Intercept)	4.958	2,614,096.849	0.000	1.000	42	IF11	5.923	169,451.330	0.000	1.000
2	c1	− 13.760	316,306.025	0.000	1.000	43	IF12	− 1.194	76,149.602	0.000	1.000
3	c2	− 7.311	132,919.886	0.000	1.000	44	IF13	− 5.565	152,714.444	0.000	1.000
4	c3	36.275	732,299.736	0.000	1.000	45	IF14	− 0.208	362,123.438	0.000	1.000
5	c4	− 36.150	1,128,248.785	0.000	1.000	46	IF15	− 3.712	181,134.756	0.000	1.000
6	c5	20.108	405,911.146	0.000	1.000	47	IF16	− 3.440	62,879.430	0.000	1.000
7	c6	1.306	410,842.052	0.000	1.000	48	IF17	8.271	127,473.185	0.000	1.000
8	c7	9.286	1,012,667.981	0.000	1.000	49	IF18	1.822	123,147.926	0.000	1.000
9	c11	− 14.339	220,132.783	0.000	1.000	50	s1	20.368	270,903.154	0.000	1.000
10	c14	6.571	56,881.870	0.000	1.000	51	s2	− 6.681	262,377.966	0.000	1.000
11	c15	− 3.120	301,610.389	0.000	1.000	52	s3	− 25.062	625,667.169	0.000	1.000
12	c10	50.178	446,569.165	0.000	1.000	53	s4	5.002	269,981.542	0.000	1.000
13	h1	1.665	15,314.144	0.000	1.000	54	s5	27.397	1,142,072.937	0.000	1.000
14	h2	− 5.516	171,718.641	0.000	1.000	55	s6	− 8.394	336,651.212	0.000	1.000
15	h3	4.163	661,162.289	0.000	1.000	56	ln1	− 1.117	124,868.301	0.000	1.000
16	h4	− 36.841	701,308.424	0.000	1.000	57	ln2	1.298	313,484.895	0.000	1.000
17	h5	14.753	239,071.266	0.000	1.000	58	ln3	− 2.481	228,680.680	0.000	1.000
18	h6	− 58.114	1,319,993.972	0.000	1.000	59	ln4	− 10.480	245,023.177	0.000	1.000
19	h7	8.613	121,302.676	0.000	1.000	60	ln5	14.163	357,466.385	0.000	1.000
20	h8	− 0.468	37,495.368	0.000	1.000	61	ln6	− 18.582	628,779.715	0.000	1.000
21	h9	13.973	420,234.026	0.000	1.000	62	ln7	9.842	525,114.283	0.000	1.000
22	h11	11.427	686,649.379	0.000	1.000	63	ln8	− 14.108	487,970.124	0.000	1.000
23	h12	− 1.232	720,857.977	0.000	1.000	64	ln9	15.025	1,307,033.796	0.000	1.000
24	h13	− 16.286	310,977.756	0.000	1.000	65	ln10	− 1.569	1,284,211.390	0.000	1.000
25	h14	21.364	640,443.859	0.000	1.000	66	StartK	0.000	5.193	0.000	1.000

Table 17 (continued)

No.	Full model									
	Coefficients	Estimate	Std. error	t value	Pr(> |t|)	No	Coefficients	Estimate	Std. error	t value	Pr(> |t|)
26	h15	1.020	446,599.076	0.000	1.000	67	CurrK	0.000	0.777	0.000	1.000
27	f1	0.000	2.654	0.000	1.000	68	SourceK	− 11.250	597,320.880	0.000	1.000
28	f2	22.664	416,770.299	0.000	1.000	69	MSEs	60.628	267,055.236	0.000	1.000
29	f3	8.189	1,019,389.766	0.000	1.000	70	Category	− 23.094	135,671.823	0.000	1.000
30	f4	31.613	441,621.077	0.000	1.000	71	Grouping	− 23.002	810,953.991	0.000	1.000
31	f5	23.491	375,903.799	0.000	1.000	72	Emp_0	− 0.677	15,277.651	0.000	1.000
32	IF1	− 5.836	117,205.548	0.000	1.000	73	X15.29	10.224	433,665.460	0.000	1.000
33	IF2	8.850	158,957.012	0.000	1.000	74	X30.49	3.975	407,546.622	0.000	1.000
34	IF3	3.289	266,299.063	0.000	1.000	75	X50.65	− 10.213	323,110.764	0.000	1.000
35	IF4	5.337	143,498.185	0.000	1.000	76	X.65	24.358	1,146,268.145	0.000	1.000
36	IF5	− 1.573	72,458.888	0.000	1.000	77	ed0	− 16.479	730,350.071	0.000	1.000
37	IF6	− 2.078	160,401.583	0.000	1.000	78	ed1	− 1.472	455,101.645	0.000	1.000
38	IF7	− 7.758	233,028.989	0.000	1.000	79	ed2	− 0.455	65,937.820	0.000	1.000
39	IF8	4.498	223,800.980	0.000	1.000	80	ed3	− 5.548	285,216.639	0.000	1.000
40	IF9	− 6.862	283,967.054	0.000	1.000	81	emp_Male	8.635	432,755.143	0.000	1.000
41	IF10	− 3.424	190,939.853	0.000	1.000						
AIC			162.000								

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 July 2018 Accepted: 27 October 2018
Published online: 18 February 2019

References

1. Wubetie HT. Missing data management and statistical measurement of socio-economic status: application of big data. J Big Data. 2017;4(1):47.
2. Johnson RA, Wichern DW. Applied multivariate statistical analysis, vol. 4. 5th ed. NJ: Prentice hall Englewood Cliffs; 2002.
3. Micro and small enterprise development policy and strategy. Ministry of Urban Development and Housing (MoUDH), 2nd edition, Addis Ababa. 2012.
4. Bosma N, van Praag M, de Wit G. Determinants of successful entrepreneurship. Research report 0002/E.2000.
5. Coduras A, Saiz-Alvarez JM, Ruiz J. Measuring readiness for entrepreneurship: an information tool proposal. J Innovat Knowl. 2016;1:99–108.
6. The small, medium and micro enterprise sector of South Africa. Commissioned by the small enterprise development agency. Research Note 2016.
7. Assefa B, Zerfu A, Tekle B. Identifying key success factors and constraints in Ethiopia's MSE development: an exploratory research. Addis Ababa: Ethiopian Development Research Institute; 2014.
8. Cochran WG. Sampling techniques. 3rd ed. New York: Wiley; 1977.
9. Lynch R, Jin Z. Exploring the institutional perspective on international business expansion: towards a more detailed conceptual framework. J Innovat Knowl. 2016;1:117–24.
10. Alves H, Ferreira JJ, Fernandes CI. Customer's operant resources effects on co-creation activities. J Innovat Knowl. 2016;1:69–80.
11. Ozkan-Canbolat E, Beraha A. Configuration and innovation related network topology. J Innovat Knowl. 2016;1:91–8.
12. Pavel R. Social entrepreneurship and vulnerable groups. J Commun Posit Pract. 2011;2:59–77.
13. Federal Democratic Republic of Ethiopia Ministry of Trade and Industry. Micro and small enterprises development strategy. Addis Ababa: Federal Democratic Republic of Ethiopia Ministry of Trade and Industry; 1997.
14. Federal Micro and Small Enterprises Development Agency. FEMSEDA annual report. 2011.
15. Federal Micro and Small Enterprises development Agency. Micro and small enterprises development strategy, provision framework and methods of implementation. Addis Ababa: Federal Micro and Small Enterprises development Agency; 2011.
16. Federal Micro and Small Enterprises development Agency. Competitive business formation guideline. Addis Ababa: Federal Micro and Small Enterprises development Agency; 2012.
17. Government of the Federal Democratic Republic of Ethiopia. Micro and small enterprise development strategy, provision framework and methods of implementation. Addis Ababa: Government of the Federal Democratic Republic of Ethiopia; 2011.
18. The Federal Democratic Republic of Ethiopia Central Statistical Agency. Statistical report on the 2013 national labour force survey. Addis Ababa: The Federal Democratic Republic of Ethiopia Central Statistical Agency; 2014.
19. Klapper L, Amit R, Guillén MF. Entrepreneurship and firm formation across countries. In: Lerner J, Schoar A, editors. International differences in entrepreneurship. National Bureau of Economic Research Conference Report. Chicago: University of Chicago Press; 2010.
20. Wang Y, Li Y, Cao H, Xiong M, Shugart Yin Yao, Jin Li. Efficient test for nonlinear dependence of two continuous variables. BMC Bioinform. 2015;16:260. https://doi.org/10.1186/s12859-015-0697-7.
21. Reshef D, Reshef Y, Finucane H, Grossman S, McVean G, Turnbaugh P, Lander E, Mitzenmacher M, Sabeti P. Detecting novel associations in large datasets. Science. 2011;334:6062.
22. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
23. Faraway JJ. Practical regression and anova using R. 2002.
24. Tibshirani R. Regression shrinkage and selection via the lasso. Toronto: University of Toronto; 1994.
25. Chattefuee S, Hadi AS. Regression analysis by example. 4th ed. New York: Wiley; 2006.
26. Agresti Alan. Categorical data analysis. 3rd ed. New York: Wiley; 2013.
27. Tsai C-F, Chen M-Y. Variable selection by association rules for customer churn prediction of multimedia on demand. London: Elsevier; 2009. p. 0957–4174. https://doi.org/10.1016/j.eswa.2009.06.07.