Austrian Journal of Statistics
July 2021, Volume 50, 32–40.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i3.1013

Comparative Performance of Three Methods to Classify Smokers Data

Amenah AL-Najafi
University of Szeged

Abstract

Since recently tobacco epidemic is one of the most important health hazards that face Iraqi individuals and communities in spite of the large information supported by the Iraqi Ministry of Health and the available statistics that link smoking with many life threatening illnesses to human. Tobacco consumption rates are increasing nowadays among university students. Iraqi Ministry of Health confirmed the need to take a serious action to support research that examines the tobacco epidemic among students, in an attempt to find the causes and the appropriate solutions. It is, therefore, our main objective is to investigate the student smokers from the University of Kufa in Iraq. The research attempted to study the behaviour of smokers using questionnaires. The performance of Latent Classes (LC) is evaluated by attempting to classify the student smokers and then compared it to two clustering methods namely K-means and Two-Step method.

Keywords: latent classes, k-means, two-step, latent gold.

1. Introduction

Smoking is a major public health problem and cigarette smoking is the single greatest cause of illness and death all around the world. The World Health Organization's (WHO) report on the state of tobacco use around the world offers terrible predictions about our future. Based on current trends in tobacco use worldwide, they tell us that we are poised on the brink of a global tobacco epidemic that could claim as many as one billion lives this century.
Previous research showed that the prevalence of smoking among university students in north Jordan was 34.0% (Khader and Alsadi 2008). While the prevalence of students smokers was 28.1% among King Faisal University in Saudi Arabia (Al Mohamed and Amin 2010). A study examined the effect of smoking on the population of Badosh District; one third of the smokers had smoking related diseases mainly respiratory disorders, most of the smokers had stressful conditions, which were the trigger of starting smoking (Alnimaa, Ahmed, and Altaiee 2008). Described the prevalence of cigarette smoking attitude among paramedical students, smoking was more common in males than females. The main sources for initiating smoking habit were friends, parents, and media.(Juni 2012), examined the prevalence of smoking among university students of medical and literature colleges. The prevalence of smoking was (48,60%) for cigarettes, (64,46%) for (sheesha). Friends were the main source of the first cigarette, followed by parents.

The LC is used as the most appropriate method to classify the data of this research. It is, therefore, the main aim of the study is to investigate the ability of classification of the LC method compared to some traditional methods. The average Silhouette index will be used to study the separation distance between the resulting clusters. Latent Gold 5.0 (Vermunt and Magidson 2005), and MATLAB are used to analyze the data.

2. Materials and methods

2.1. Latent classes model

Let $\mathbf{x}$ denote a response vector of p binary manifest variables. The general form of the marginal distribution of the manifest variables is defined as follows:

$$f(\mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1-\pi_{ij})^{1-x_i}, \tag{1}$$

where π_{ij} $(i = 1, \ldots, p; j = 0, \ldots, K-1)$ denotes the conditional positive response probabilities and η_j is the prior probability that a randomly chosen individual is in class j $(\sum_{j=0}^{K-1} \eta_j = 1)$. The posterior probability of an individual in the response vector $\mathbf{x}$ belongs to category j, can be formulated as follows:

$$h(j|\mathbf{x}) = \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1-\pi_{ij})^{1-x_i} / f(\mathbf{x}).$$

The parameters of Latent Classes model (LC) are typically estimated by means of maximum likelihood (ML) from the equation 1, the E-M algorithm is used to derive the ML estimates, see (Moustaki 1996) for more details. The maximum likelihood estimates are:

$$\hat{\eta_j} = \sum_{h=1}^{n} h(j|\mathbf{x}_h)/n \quad (j = 0, 1, \cdots, K-1), \tag{2}$$

and

$$\hat{\pi_{ij}} = \sum_{h=1}^{n} x_{ih} h(j|\mathbf{x}_h)/(n\hat{\eta_j}) \quad (i = 1, \cdots, p; j = 0, 1, \cdots, K-1), \tag{3}$$

2.2. K-mean

K-means algorithm suggested by (MacQueen 1967), this algorithm is unsupervised that usually use in clustering observations into a specific number of disjoint clusters having the nearest centroid. The Common distance measure to the assigned nearest centroid is the Euclidean distance.

Let the dataset of N data points be $\{x_1, \ldots, x_N\}$ such that each x belong to R^d where d is the number of dimensions.The k-means algorithm partitions the given data into k clusters, in order to illustrate the of the K-means clustering method to the initial choice of cluster main point: Firstly the data points is selected to be the initial partitioned into K clusters randomly. Secondly, for each sample, the distance is calculated from the observation to the centroid of the cluster; if the sample is closest to its own cluster then leave it else select another cluster. Thirdly, steps 1 and 2 are repeated until no observations are moved from one cluster to another. Now when step 3 ends the clusters are fixed and each sample is allocated a cluster, which results in the lowest possible distance to the centroid of the cluster. To find K cluster centroids the (mean squared error, MSE) is used such that:

$$1/N \sum_{i=1}^{N} [\min_{j} d^2(x_i - m_j)] \text{ is minimized,}$$

where $d^2(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j and m_j cluster sentroids.

2.3. Two-step

The technique was developed by Chiu and et al (2001) for the analysis of large data sets. The technique can handle scale and ordinal data in the same model and can routinely determine the ideal number of clusters. The technique recognises groupings by running pre-clustering first and then clustering.

In pre-clustering, the technique uses a Clustering Feature (CF) for clustering. It scans the data records one by one and decides if the existing record should be combined with the previously formed clusters or starts a new cluster, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. The method used Euclidian and Log-Likelihood distance. The CF tree consists of levels of nodes, each node having a number of entries. A leaf entry is a final sub-cluster. For each record, starting from the root node, the nearest child node is found recursively, descending along the CF tree. Once reaching a leaf node, the algorithm finds the nearest leaf entry in the leaf node. If the record is within a threshold distance of the nearest leaf entry, then the record is added into the leaf entry and the CF tree is updated. Otherwise, it creates a new value for the leaf node. If there is enough space in the leaf node to add another value that leaf is divided into two values and these values are distributed to one of the two leaves, using the farthest pair as seeds and redistributing the remaining values based on the closeness criterion Chiu and et al (2001).

After the CF tree is rebuilt, the procedure checks if these values can fit in the tree without increasing the tree size. The values that do not fit anywhere are considered outliers.

The clustering stage has sub-clusters resulting from the pre-cluster step as input and groups them into the desired number of clusters (Bacher and Vogler 2004). The log-likelihood distance is used to handle mixed-type attributes. Let k^A total number of continuous variables used in the procedure, k^B is the number of categorical variables employed in total, N_i number of data records in cluster i, and $< i, s >$ is the index that represents the cluster formed by combining clusters i and s, the log-likelihood distance between two clusters i and s is defined as:

$$d(i, s) = \xi_i + \xi_s - \xi_{<i,s>},$$

where

$$\xi_i = -N_i[\sum_{k=1}^{k^A} \frac{\log(\sigma_k^2 + \sigma_{ik}^2)}{2} + \sum_{k=1}^{k^B} E_{ik}]$$

$$E_{ik} = -\sum_{k=1}^{L_k} \frac{N_{ikl}}{N_i} \log \frac{N_{ikl}}{Ni},$$

where σ_k^2 the estimated variance of the $k - th$ continuous variable in whole data, σ_{ik}^2 is the variance of the $k - th$ continuous variable in cluster i, L_k number of categories for the $k - th$ categorical variable and N_{ikl} is the number of data records in cluster i whose $k - th$ categorical variable takes the $l - th$ category.

The number of clusters can be determined using Schwarz's Bayesian Criterion (BIC) and the Akaike Information Criterion (AIC) as the clustering criterion.

3. Description of data

A multistage sampling method is applied. A random sample of Faculties from the University of Kufa is first selected, they were Art, Jurisprudence, Administration and Economics, Engineering, Physical Education, Medicine, the sample included male students only due to

the traditions of the community at the University of Kufa. The selected universities halls are used to provide guidance on the objectives of the questionnaire, to ensure the confidentiality of the questionnaire; numbers replaces the names. Only a list of smokers is obtained.
In the second stage, simple random sampling is used to select the sample from students who smoked only from each faculty. After explaining the method of filling the questionnaire, a total of 160 structured questionnaires are distributed and 150 (94%) are fully completed. A total of 10 questionnaires are returned unfilled.
The questionnaire comprised a mix of open-ended and multiple-choice questions, aimed at collecting data on the student's demographic characteristics and their smoking behaviour. The questionnaire includes additional information on current and previous smoking status, age of initiation, and reasons for starting smoking. Questions related to reasons for starting smoking are open-ended. The smoking impact of the student, on their family members and closest friends, is obtained. Questions 1 to 7 related to addiction are illustrated in Table 1.

Table 1: Questionnaires included in the analysis

Question Label	Questionnaires
Q1	How old were you when you began smoking?
Q2	What is the reason that encouraged you to smoke?
Q3	Is your smoking have an impact on your relatives or friends?
Q4	From the moment you wake up how long it takes before you light up the first cigarette?
Q5	Any cigarette can do without, the first one of the day or otherwise
Q6	Do you smoke more in the morning than the rest of the day?
Q7	Do you smoke when you are sick or if you have to stay in bed?

A total of 7 structured questionnaires are distributed to student's age 18-23 year.
Question one is categorised into two categories (10-19), and (19-30). Question two is categorised into two categories namely psychological reasons, and social reasons. Question three, five, six, and seven are categorised as (No), and (yes). Question four is categorised into two categories; less than an hour, and greater than an hour.

4. Results

The majority of the smokers are age 10-19, 103 (68.7%). About, 92 (61.3%) of them have grades below seventy in the previous year. They smoke 25 cigarette and over per day, and spend on tobacco at least 30,000 Iraqi Dinar a month. Almost the entire student is aware of the dangers of smoking, 40 (93.3%), and out of 150, 105 (70%) admits that they know the effect of smoke on their health.

Table 2 shows the proportion of student smokers within each category for each question. The highest percentages is among age 10 and 19 years old, 68.7% of smokers cannot do without the first cigarette of the day, 68.7% of smokers smoke less in the morning compared to the rest of the day, and 59.3% smokers light up the first cigarette in less than an hour from the moment they wake up.

The Latent Class Analysis (LCA) is illustrated in Table 3 to 5. The model L^2 statistic is shown in Table 3 indicates the amount of the association among the variables that remained unexplained after estimating the model; the lower the value, the better the fit of the model to the data.
One criterion for determining the number of classes is to look at the 'p-value' for each model under the assumption that the null hypothesis of this test satisfies model holds true in the population. Generally, among models for which the p-value is greater than 0.05 will provides

an adequate fit, the fewest number of parameters would be selected. Using these criteria, the best model is given by Model 2, the 2-classes model, with a p-value of 0.12, and a number of parameters equal to 14.

Table 2: Descriptive statistics of survey questions

Question	Category	Number (N=150)	Percentages
Q1	10-19	103	68.7
	19-30	47	31.3
Q2	Psychological	80	53.3
	Social	70	46.7
Q3	No	82	54.6
	Yes	68	45.3
Q4	<an hour	89	59.3
	>an hour	61	40.7
Q5	No	103	68.7
	Yes	47	31.3
Q6	No	103	68.7
	Yes	47	31.3
Q7	No	87	58.0
	Yes	63	42.0

Table 3: Model summary

Model	Class	LL	$BIC(LL)$	No. of Parameter	L^2	$d.f.$	p-value
I	1	-690.1273	1414.3290	7	162.1400	120	0.0063
II	2	-673.8678	1422.8941	14	129.6210	112	0.12
III	3	-669.0204	1443.2844	23	119.9263	104	0.14
IV	4	-664.2773	1484.8842	31	112.4400	96	0.12
V	4						

The R^2 values are in the right-most column of Table 4 specifies how much of the variance of each indicator is explained by these two cluster models. For example, the model explains 7.32%, 30.74%, and 34.34% of the variance of the variables 1, 4, and 7 respectively.

Table 4: Model parameters

Models for Indicators		Class1	Class2	R^2
Q1	0	0.2996	-0.2996	0.0732
	1	-0.2996	0.2996	
Q2	0	-0.0307	0.0307	0.0008
	1	0.0307	-0.0307	
Q3	0	-0.0271	0.0271	0.0007
	1	0.0271	-0.0271	
Q4	0	0.6431	-0.6431	0.3074
	1	-0.6431	0.6431	
Q5	0	0.3263	-0.3263	0.0864
	1	-0.3263	0.3263	
Q6	0	-0.1394	0.1394	0.0138
	1	0.1394	-0.1394	
Q7	0	-1.1099	1.1099	0.3434
	1	1.1099	-1.1099	

Table 5 contains aggregated class membership probabilities for ranges of values of indicators. The first row of the table contains the overall probability of being in a class, the size of

Table 5: Profile and probability means for 2-classess

Indicators		Profile		Probability Means	
		Class1	Class2	Class1	Class2
Overall		0.6471	0.3429	0.6471	0.3429
Q1	0	0.7773	0.4130	0.7446	0.2444
	1	0.2227	0.4870	0.4646	0.4344
Q2	0	0.4229	0.4434	0.6441	0.3449
	1	0.4771	0.4466	0.6719	0.3281
Q3	0	0.4374	0.4643	0.6460	0.3440
	1	0.4624	0.4347	0.6704	0.3294
Q4	0	0.7901	0.2163	0.8770	0.1230
	1	0.2099	0.7837	0.3371	0.6629
Q5	0	0.7842	0.4978	0.7422	0.2478
	1	0.2148	0.4022	0.4489	0.4411
Q6	0	0.6473	0.7622	0.6191	0.3809
	1	0.3427	0.2378	0.7404	0.2494
Q7	0	0.3711	0.9804	0.4187	0.4813
	1	0.6289	0.0196	0.9872	0.0128

each class is also reported, conditional probabilities associated with each category of nominal indicator variables (these probabilities sum to 100% across rows).
Class one that contains around 65% of the cases, class two contains 34%. The conditional probabilities show the differences in response patterns that distinguish the class one from two. For instance, smokers in class one are much more likely lights up their first cigarette in less than an hour (0.877) after they wake up, (0.784) could not do without their morning first cigarette, (0.647) smoke mainly in the morning than at any other time, and (0.987) of smokers continue to smoke even when they are sick in bed. Figure 1 demonstrates Uni-plot that confirm the distribution of student's smoker over two classes.

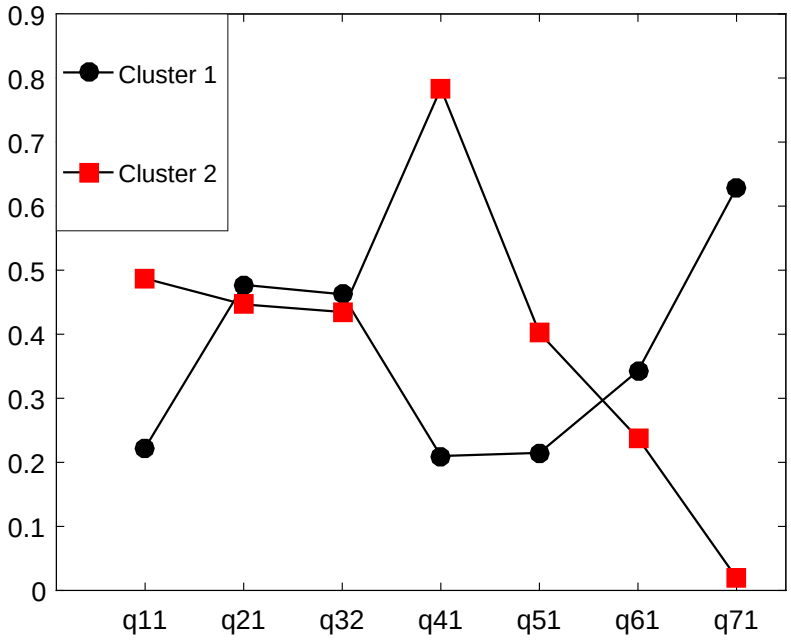

Figure 1: Uni-plot for two classes

Table 6 shows the patient's proportion distribution over two classes for three classes techniques. Details of the analysis of the results for each method are explained below.

The LCA reveals two distinct classes. The most prevalent classes (65.71%) primarily contain smokers age 10-19 (80.2%) and nearly (59.3.1%) of them smoke for psychological reasons, and (40.7%) for social reasons. About (51.9%) of smokers have no impact on their relatives or friends. Smokers in this category light up their first cigarette in less than an hour (100%) after they wake up. They could not do without the first cigarette of the day (81.5%) of them, and smoked less in the morning (65.4%), in addition to that, they do not smoke when they were sick (56.8%). The second class (34.29%) primarily contains smokers (55.1%) age 10-19 years. Around (58%) of smokers have no impact on their relative or friends. About (88.4%) light up their first cigarette in an hour or more, as well as they do smoke when they are sick.

K-means, the first class, which includes around (40.7%), contains smokers (55.7%) age between 10 and 19 years old. About (78.7%) of smokers have no impact on their relatives or friends. Nearly (82%) of Smokers could not do without the first cigarette of the day. The second class (59.3%) primarily includes smokers (77.5%) age 10-19, (92.1%) of them smoked less in the morning.

Two-Step method it is distinguished by the following two categories, the first class, which includes around (47.1%), contains smokers age 10-19 (51.5%), about (57.6%) smoke for psychological reason. A (100.0%) of smokers do smoke when they are sick. The second class (52.9%) primarily includes smokers (82.1%) age 10-19, nearly (50.0%) for a social and for psychological. About (84.5%) light up their first cigarette in less than an hour, and (65.5%) of them smoked less in the morning as well as (75.0%) they do not smoke when they are sick.

Table 6: Classification of three clustering techniques

Category		Latent Class		K-Mean		Two Step	
		Class No.		Class No.		Class No.	
		1	2	1	2	1	2
Overall		99(65.71%)	51(34.29%)	61(40.7%)	89(59.3%)	33(47.1%)	37(52.9%)
Q1	0	65(80.2)	38(55.1%)	34(55.7%)	69(77.5%)	34(51.5%)	69(82.1%)
	1	16(19.8%)	31(44.9%)	27(44.3%)	20(22.5%)	32(48.5%)	15(17.9%)
Q2	0	48(59.3%)	32(46.4%)	29(47.5%)	51(57.3%)	38(57.6%)	42(50.0%)
	1	33(40.7%)	37(53.6%)	32(52.5%)	38(42.7%)	28(42.4%)	42(50.0%)
Q3	0	42(51.9%)	40(58.0%)	48(78.7%)	34(38.2%)	36(54.5%)	46(54.8%)
	1	39(48.1%)	29(42.0%)	13(21.3%)	55(61.8%)	30(45.5%)	38(45.2%)
Q4	0	81(100.0)	8(11.6%)	24(39.3%)	65(73.0%)	18(27.3%)	71(84.5%)
	1	0 (0.0%)	61(88.4%)	37(60.7%)	24(27.0%)	48(72.7%)	13(15.5%)
Q5	0	66(81.5%)	37(53.6%)	50(82.0%)	53(59.6%)	33(50.0%)	70(83.3%)
	1	15(18.5%)	32(46.4%)	11(18.0%)	36(40.4%)	33(50.0%)	14(16.7%)
Q6	0	53(65.4)	50(72.5%)	21(34.4%)	82(92.1%)	48(72.7%)	55(65.5%)
	1	28(34.6%)	19(27.5%)	40(65.6%)	7(7.9%)	18(27.3%)	29(34.5%)
Q7	0	35(43.2%)	52(75.4%)	42(68.9%)	45(50.6%)	66(100.0%)	21(25.0%)
	1	46(56.8%)	17(24.6%)	19(31.1%)	44(49.4%)	0 (0.0%)	63(75.0%)

The average Silhouette index refers to a method of interpretation and validation of consistency within clusters of data. The technique estimates average Silhouette index for each cluster and overall average silhouette index. The average silhouette index is used to measures how similar a point is to its cluster versus the next closest cluster. This is a ratio of the distances to the cluster centres, normalized so that "1" is a perfect match to its cluster and "-1" a perfect mismatch. Table 7 shows the average silhouette of different clusters. The results shows, when number of clusters is two, the average Silhouette indexes are 0.222, 0.220, and 0.198 for the Latent Class, the Two-Step and the K-means respectively.

Table 7: Average silhouette index

Average Silhouette			
Number of Clusters	Latent Class	Two Step	K-Mean
2	**0.222**	0.220	0.198
3	0.20	0.195	0.193
4	0.106	0.160	0.187

5. Conclusion

The main characteristic of the study sample are nearly 70% of the smokers are among age 10 to 19 years old, and about, 93% of them have low university's grades in the previous year. The main aim of this study is to evaluate the performance of Latent Class to classify the smokers and then compare its results with two clustering methods namely K-means and Two-Step, as well as determining the optimum category to be used for building the model.

Five categories are chosen to compare each method. The results are fairly similar. Hence, three categories are chosen instead, but the results show that the optimum category is two for all methods. The Latent Class method give the best results using two categories, followed by a Two-step method, and then the K-means.

It is not surprising that the question seven (smokers light up the first cigarette in less than an hour from the moment they wake up) and question four (smokers do smoke when they are sick) dominates cluster one, which represent very heavy smokers (addict) using LC and Two Step method. While question five (Smokers could not do without the first cigarette of the day) and question six (smoker smoke in the morning more than the rest of the day) dominates cluster two, which denote moderate smokers using the K-means. However, the second question (The reason encourage the smoker to smoke), and third question (smoker smoking have an impact on relative and friends) are the least effective.

The Average Silhouette index is used to evaluate the performance of the three methods and find out which one is most appropriate for our data. The results indicate that the sample is well cluster using latent class compared to Two Step and K-means. However, when the number of cluster equal to two, the average silhouette index for Latent Class is 0.222, which is the best result, followed by the Two-Step and the K-means is the last. It is interesting to note that the Silhouette index when number of clusters three or four for Latent class is less than when the number of clusters is two. This is also true for the Two-Step and the K-means methods. In conclusion, Latent class indicates that smokers are more matched to its own cluster than The Two-Step method and the K-means.

In summary, based on our sample that is taken from the University of Kufa, the analysis emerged two distinct classes with the following characteristics: firstly about (66%) heavy smoker students that may be addicted to nicotine, secondly around (34%) lighter smokers students. However, previous research conducted by Juliana (2015) investigated different types of student smokers of Colorado State University in the USA. The results revealed that a four classes model. The classes included addicted smokers, non-endorsing smokers, stress smokers, and social smokers (Rosa and Aloise-Young 2015). The data revealed two different classification groups for three clustering methods. The Latent Class offered the best results with our data but different clustering methods might suit other types of data. An important conclusion arising from this study as well is that the dominance of one method over others is not even.

References

Al Mohamed HI, Amin TT (2010). "Pattern and Prevalence of Smoking Among Students at King Faisal University, Al Hassa, Saudi Arabia." *Eastern Mediterranean Health Journal*, **16**, 56–64. URL https://doi.org/10.26719/2010.16.1.56.

Alnimaa BA, Ahmed MM, Altaiee WG (2008). "Tobacco Smoking Habit Among Badosh District People." *Medical Journal of Tikrit*, **1**(141), 221–227.

Bacher Jand Wenzig K, Vogler M (2004). "SPSS TwoStep Cluster-A First Evaluation."

Chiu T, et al (2001). "A robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment." *In Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 263–268.

Juni FH (2012). "Cigarette Smoking Habits Among Paramedical Students in Baghdad-Iraq." *Al-Qadisiah Medical Journal*, **8**(13), 188–194.

Khader YS, Alsadi AA (2008). "Smoking Habits Among University Students in Jordan: Prevalence and Associated Factors."

MacQueen J (1967). "Some Methods for Classification and Analysis of Multivariate Observations." *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, **1**(14), 281–297. URL https://doi.org/10.3109/10826084.2015.1018549.

Moustaki I (1996). *Latent Variable Models for Mixed Manifest Variables*. Ph.D. thesis, The London School of Economics and Political Science (LSE).

Rosa JD, Aloise-Young P (2015). "A Qualitative Study of Smoker Identity Among College Student Smokers." *Substance use & misuse*, **50**(12), 1510–1517. URL https://doi.org/10.3109/10826084.2015.1018549.

Vermunt JK, Magidson J (2005). "Latent GOLD® Choice 4.0 User's Manual." *Statistical Innovations Inc., Belmont, MA*.

Affiliation:

Amenah Al-Najafi
Szeged University
Telephone: +36/702/224179
E-mail: ma_rw114@yahoo.com