Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2022 (8 Oct);36.116. https://doi.org/10.47176/mjiri.36.116

A Cluster-wise Linear Regression Model to Investigate the Effect of Demographical and Clinical Variables on the Average Depression Score

Zahra Zamaninasab[1], Hamid Najafipour[2], Moghaddameh Mirzaee[1]*, Abbas Bahrampour[1]

Received: 7 Nov 2021 Published: 8 Oct 2022

Abstract

Background: Depression is a prevalent illness in the world. Given the importance of mental disorders, many researchers have investigated the effects of different variables on average depression scores. In this study, we decided to investigate the effect of some explanatory variables on the average depression score.

Methods: The data were provided from the second phase of the Kerman Coronary Artery Diseases Risk Factors study (KERCADRS), which took place between 2014 and 2018. To obtain more precise connections between depression ratings and predictor variables, we employed a cluster-wise linear regression model.

Results: The total number of the participants in this study was 9811, out of whom 2144 were allocated to cluster 1, 4540 to cluster 2, and 3127 to cluster 3. The average depression score was 13.76 ± 7.6 in cluster 1, 4.39 ± 4.7 in cluster 2, and 10.83 ± 6.7 in cluster 3. However, the average depression score for all the data was 8.5 ± 7.2. In all the clusters, the average depression score of females was significantly greater than that of men ($p < 0.001$). In cluster 1, the age category of 35-54 years, in cluster 2, the age category of 55-80 years, and in cluster 3, the age category of 15-34 years had a maximum average depression score.

Conclusion: We may classify the 3 clusters as having a low (cluster 2), moderate (cluster 3), or high (cluster 1) depression score, according to the age group with the highest artery diseases risk. The patients were 55-80 years, 15-34 years, and 35-54 years in cluster 2 (low), cluster 3 (moderate), and cluster 1 (high), respectively.

Keywords: Depression, Coronary Artery Diseases, Clustering

Conflicts of Interest: None declared
Funding: None

Cite this article as: Zamaninasab Z, Najafipour H, Mirzaee M, Bahrampour A. A Cluster-wise Linear Regression Model to Investigate the Effect of Demographical and Clinical Variables on the Average Depression Score. *Med J Islam Repub Iran*. 2022 (8 Oct);36:116. https://doi.org/10.47176/mjiri.36.116

Introduction

Depression is a prevalent condition worldwide, affecting around 264 million people (1). Depre ssion is distinct from normal mood swings and brief emotional reactions to ordinary difficulties. It is the main cause of disability globally and a significant contribution to the global illness burden (2). Women are more likely than males to suffer from depression, and it is the primary cause of the disease's burden among women (3). In general, between 10% and 25% of women and 5% to 10% of men experience depression at some point in their lives (4). Given the se-

Corresponding author: Dr Moghaddameh Mirzaee, M_mirzaee@kmu.ac.ir

1. Modeling in Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
2. Department of Physiology and Pharmacology, School of Medicine, Physiology Research Center, Kerman University of Medical Sciences, Kerman, Iran

↑What is "already known" in this topic:
In the analysis of heterogeneous data, especially in big data, fitting only 1 model for the whole data set does not have enough accuracy and may lead to a model with weak goodness of fit and low accuracy.

→What this article adds:
The key benefit of this work is that we employed the Cluster-wise Linear Regression (CLR) approach on a massive data frame to execute a distinct regression model in each cluster rather than applying a single model to the entire data. With more precise models, it would be feasible to determine the best treatment for the patients in each cluster.

[DOI: 10.47176/mjiri.36.116]

verity of mental problems, mental health has gained more attention in Iran in recent years (5). New efforts have been initiated by health institutions and clinics to enhance the detection and treatment of depression (6). Additionally, the frequency of mental diseases among individuals in Iran's general population was investigated. According to Noorbala et al's 1999 National Initiative on Health and Illness (7), the prevalence of mental illnesses was assessed to be 21% (25.9% among women and 14.9% among men). Mohammadi et al did another nationwide study and assessed the prevalence of mental diseases to be 17.1% (23.4% for women and 10.8% for men) (8). Cardiovascular patients, compared with healthy persons, have increased levels of anxiety, sadness, and perceived stress (9-11). Psychological variables have a significant influence in the genesis of cardiovascular disease. Thus, treating psychological illnesses might significantly aid in the prevention of these diseases (9, 10). By providing adequate mental health services across the nation, the government can help reduce the prevalence of depression and anxiety disorders, as well as various physical ailments associated with these disorders, most notably coronary artery disease. Prudent planning for the provision of mental health care requires precise data on the prevalence of these diseases in the nation. The purpose of this work was to develop a statistical model of the average depression score (ADS) using a variety of demographic and clinical data. While multiple linear modeling and logistic regression are regularly used methodologies, we employed Cluster-wise Linear Regression (CLR) to analyze the connection between the response variable and predictors in this study (12). Clustering techniques are often used for 2 reasons: to analyze data and to increase the predicted accuracy of models. Second, persistent clusters with consistent nonessential changes might be utilized to classify, target, and interpret observed patients (13). We could cluster observations in clusters and fit the regression models in each cluster in the way that the mean squared error (MSE) of each cluster-specific regression model reach the minimum value. As a result, the accuracy of each cluster model was at its maximum level. This method could be also utilized to reduce the negative effect of heterogeneity in data, especially in the big data concept (14, 15). The positive aspect utilizing the CLR method is that we can group the data into the clusters that have their model and the people in each cluster are interpreted with their specific cluster model. This strategy avoids using a single model to interpret all of the data. For example, a variable may not be important when we fit one model to all of the data, but it may be significant in some or all clusters when we use the CLR technique.

Methods

Study Population

The data for this study were derived from the second phase of the Kerman Coronary Artery Diseases Risk Factors study (KERCADRS), which was conducted in 2014-2018 in Kerman province, the largest province in southeast of Iran. The Ethics Committee of Kerman University of Medical Sciences accepted the research protocols (Ethical code: IR.KMU.REC.1392.405). The research population comprised of 10,015 individuals aged 15 to 80 years who were Kerman residents and were recruited through the cluster sampling technique. After removing the missing data, our research included 9811 individuals. Demographic data were gathered. The whole project's methodology has been published in the Iranian journal of public health (16).

Instruments

Beck's Depression and Anxiety Scales were used to examine depression and anxiety symptoms in this study (17). Both surveys have also been verified in Farsi (18). The Beck depression index (BDI) is a 21-item questionnaire used to examine depressive symptoms. Each item is graded on a scale of 0 to 3. The BDI score is calculated as the sum of the values assigned to each item (total: 63). Scores of 1 to 10 are regarded normal, 11 to 16 indicate mild mood disturbance, 17 to 20 indicate borderline clinical depression, 21 to 30 indicate moderate depression, 31 to 40 indicate severe depression, and above 40 indicate serious depression. Additionally, the Beck anxiety index (BAI) scale is a questionnaire used to examine anxiety symptoms, consisting of 21 items. Each question is graded on a scale of 0 to 3. The BAI total score is calculated as the sum of the values assigned to each question (total: 63). Low or normal anxiety is defined as a score of 0 to 21, moderate anxiety is defined as a score of 22 to 35, and perhaps worrying anxiety is defined as a score of 36 or above (19). The analytic guide for the Global Physical Activity Questionnaire (GPAQ) was used to measure physical activity.

Variables

In this study, the continuous form of depression was considered as the response variable.

We used the predictor variables, involving age (continuous form), sex (male, female), marital status (single, married, divorced, and widowed), physical activity (low, medium, and high), current cigarette smoking (yes, no), current tobacco smoking (yes, no), diabetes (normal, prediabetic, and diabetic), hypertriglyceridemia (yes, no), hypercholesterolemia (yes, no), anxiety (continuous form), and body mass index (BMI) (underweight, normal, overweight, and obese).

In diabetes variable, a prediabetic person is someone with fasting blood sugar (on) of between 100 to 126 mg/dL and a diabetic patients is one with FBS over 126 mg/dL.

Patients with hypertriglyceridemia have triglycerides ≥150 mg/dL or are take medication to treat elevated triglycerides, whereas those with hypercholesterolemia have total cholesterol ≥240 mg/dL or take medication to treat elevated cholesterol. Furthermore, in the BMI variable, the values of BMI ≤18.5 refer to underweight people, 18 to 25 BMI values refer to normal people, 25 to 30 BMI value refer to overweight people, and the values of BMI ≥30 indicate obese people.

[DOI: 10.47176/mjiri.36.116]

Statistical Analysis

In this paper, we aimed to statistically model some demographic and clinical variables on the average depression score. Therefore, a multiple linear regression model was primarily fitted on the whole data and the R-squared and the mean squared error (MSE) of this model were recorded. Following that, the cluster-wise regression was employed. This approach is based on a combination of 2 methods, clustering and regression, and it discovers an ideal partition of data (k clusters) and regression functions inside clusters with the lowest error and highest goodness of fit. It might be used to identify patterns in data when several patterns are likely to exist. Herein, we presented a brief description of the CLR method steps: 1) The data were allocated to k clusters randomly; in each cluster, a multiple linear model was then fitted. 2) The MSE for each cluster- specific regression model was calculated (we now have k MSEs). 3) In step 3, each individual was reallocated to the cluster with the minimum MSE. 4) A multiple linear regression was fitted on each new cluster and the MSE of each cluster was computed and the average of these *k* MSEs was the overall MSE of each step. 5) Steps 2 to 4 were repeated until the overall MSE did not change. When the loop was finished, we obtained the minimum overall MSE. Accordingly, to specify the true number of clusters (k) for our data, the CLR model was fitted for different number of clusters: k = 3, 4, 5, 6, 7, and 8. We should choose the value k, which reflects the maximum separation between groups, because all clustering algorithms are anticipated to increase similarity within clusters while minimizing similarity across clusters. To achieve this, the standard deviation (SD) of MSEs of the clusters for each k was computed (for example, if K = 3 the SD of (MSE cluster1, MSE cluster2, MSE cluster3) was calculated, and in a similar fashion for other values of k and the largest SD value was picked. Then, the corresponding k was chosen as the true number of clusters. In present study, k = 3 clusters resulted in maximum SD among all number of k; thus, the true number of clusters that make the best separation in the data was decided as 3. All analyses were done in the R programming language Version 4.0.1 (20, 21).

Results

Table 1 depicts the baseline demographic and clinical characteristics both totally and by clusters. The total number of the patients in this study was 9811, out of whom 2144 were allocated to cluster 1, 4540 to cluster 2, and 3127 to cluster 3. The overall mean age was 45.8 ± 15.2 years and 59.7% of the patients were women. The percentages shown in Table 1 are the cluster-specific percentages; for example, in cluster 1, 50% of the patients

Table 1. Baseline demographic and clinical characteristics by cluster and totally

Characteristics	Cluster 1	Cluster 2	Cluster 3	Total
Mean ±SD				
Age	47.5 ± 16.0	43.9 ± 15.4	47.4 ± 14.1	45.8 ± 15.2
Anxiety	9.4 ± 7.9	7.1 ± 7.2	8.7 ± 7.9	8.1 ± 7.7
N (%)				
Sex, female	1324 (61.7%)	2671 (58.8%)	1859 (59.4%)	5854 (59.7%)
Marital Status				
Single	369 (17.25)	702 (15.5%)	142 (4.5%)	1213 (12.4%)
Married	1574 (73.4%)	3580 (78.8%)	2857 (91.4%)	8011 (81.6%)
Divorced	35 (1.6%)	51 (1.1%)	13 (0.4%)	99 (1%)
Widowed	166 (7.7%)	207 (4.6%)	115 (3.7%)	488 (5%)
Physical Activity				
Low	1050 (50%)	2135 (47%)	1481 (47.4%)	4666 (47.6%)
Medium	802 (37.4%)	1673 (36.8%)	1193 (38.1%)	3668 (37.4%)
High	292 (13.6%)	732 (16.1%)	453 (14.5%)	1477 (15%)
Cigarette Smoking, yes	189 (8.8%)	391 (8.6%)	308 (9.8%)	888 (9%)
Tobacco Smoking, yes	225 (10.5%)	387 (8.5%)	241 (7.7%)	853 (8.7%)
Diabetes				
Normal	1150 (53.6%)	3465 (76.3%)	2432 (77.8%)	7047 (71.8%)
Pre-Diabetic	523 (24.4%)	957 (21.1%)	200 (6.4%)	1680 (17.1%)
Diabetic	471 (22.0%)	118 (2.6%)	495 (15.8%)	1084 (11%)
Hypertriglyceridemia, yes	782 (36.5%)	1225 (27.0%)	1073 (34.3%)	3080 (31.4%)
Hypercholesterolemia, yes	213 (9.9%)	362 (8.0%)	279 (8.9%)	854 (8.7%)
BMI				
Underweight	81 (3.8%)	191 (4.2%)	104 (3.3%)	376 (3.8%)
Normal	597 (27.8%)	1459 (32.1%)	967 (30.9%)	3023 (30.8%)
Overweight	843 (39.3%)	1767 (38.9%)	1238 (39.6%)	3848 (39.2%)
Obese	623 (29.1%)	1123 (24.7%)	818 (26.2%)	2564 (26.1%)
Total Observation (n)	2144	4540	3127	9811

[DOI: 10.47176/mjiri.36.116]

had a low level of physical activity, 37.4% a medium level of physical activity, and 13.6% a high level of physical activity. In the following. We interpreted the numbers in Table 1 in another way with the row percentage; for instance, there were 1477 patients with a high level of physical activity and among them, 732 (almost 50% of all the people with a high level of physical activity,$\frac{732}{1477}$) were allocated to cluster 2. However, 11% of diabetics were assigned to cluster 2, while 35% of those with hypertriglyceridemia were assigned to cluster 3, and so on.

The regression coefficients of the overall multiple regression model and the CLR model are represented in Table 2. In the overall multiple regression model, which was fitted on all the 9811 patients, the R-squared was 41%; this means that the variance participation accounted by the model was 41%. Furthermore, 3 variables of marital status, diabetes, and hypercholesterolemia were not significant in the overall model. In this model, for each 1-year increase in age, the ADS increased by 0.04. For each unit increase in the anxiety score, the mean score of depression increased by 0.58. In addition, the mean depression score of women was 1.19 units greater than that of men. For people with a medium level of physical activity, the ADS decreased by 0.34 in comparison with those who had a low level of physical activity. This decreasing pattern was found to be the same for people with a high level of physical activity by a 0.79 unit decrease in ADS compared to those with low physical activity levels. As shown in Table 2, after the CLR model was fitted for k = 3 clusters, the R-squared were 85%, 77%, and 85% for cluster1, cluster2, and cluster 3, respectively, which significantly increased compared to the overall model's R-squared. Furthermore, the MSE was 8.5 for cluster 1, 4.95 for cluster 2, and 6.47 for cluster 3; meanwhile, in the overall model, the MSE was 31.04. These 2 metrics (R-squared and MSE) suggested that the CLR technique could help in the fitting of more powerful models by clustering data into k clusters, each with a lower MSE than an overall model fitted for all data. In cluster 1, all the variables had a significant effect on ADS, except for tobacco smoking and physical activity. However, married people had 8.53 units greater average depression score ADS, divorced people had 5.05 units greater ADS, and widowed people had 0.84 units greater ADS in comparison with single participants. People with hypercholesterolemia also had further ADS, 0.55 units greater than people with a normal level of cholesterol. More details of regression coefficients of cluster 1 are illustrated in Table 2. For cluster 2, all the variables significantly affected ADS, except for cigarette smoking, tobacco smoking, and hypertriglyceridemia. In contrast with cluster1, married people in cluster 2 had lower ADS (0.74 units of decrease) compared to single people. Diabetic people had a huge increase in ADS (16.84 units) compared to nondiabetic people. Moreover, the ADS for prediabetic people had a 0.73-unit increase compared to nondiabetic people. Further details concerning the regression coefficients of cluster 2 are shown in Table 2. In cluster 3, all the variables had a significant effect on ADS, except for tobacco smoking, BMI, and hypercholesterolemia. In contrast to cluster 1 and cluster 2, in cluster 3, being married and widowed had a negative effect on ADS. Married people had 10.04 units of decrease in ADS compared to single people, and widowed people had 2.59 units of decrease in comparison with the single people (Table 2).

Figure 1 represents the convergence curve of the R-

Table 2. Regression coefficients for overall and CLR Model

	Cluster 1			Cluster 2			Cluster 3			Overall Model		
Variable	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
Intercept	5.86	0.45	<0.001	1.53	0.24	<0.001	15.21	0.37	<0.001	4.77	0.41	<0.001
Age	0.03	0.006	<0.001	0.02	0.002	<0.001	0.03	0.004	<0.001	0.04	0.005	<0.001
Anxiety	0.55	0.008	<0.001	0.34	0.005	<0.001	0.52	0.006	<0.001	0.58	0.007	<0.001
Sex, Female	1.60	0.16	<0.001	0.66	0.08	<0.001	1.39	0.11	<0.001	1.19	0.13	<0.001
Marital Status												
Married	8.53	0.23	<0.001	-0.74	0.11	<0.001	-10.04	0.24	<0.001	-0.30	0.20	0.141
Divorced	5.05	0.54	<0.001	1.49	0.33	<0.001	4.34	0.74	<0.001	2.01	0.59	0.001
Widowed	0.84	0.36	0.021	2.33	0.21	<0.001	-2.59	0.35	<0.001	0.27	0.35	0.443
Physical Activity												
Medium	-0.34	0.14	0.012	-0.28	0.07	<0.001	-0.21	0.10	0.031	-0.34	0.12	0.005
High	-0.37	0.20	0.064	-0.49	0.10	<0.001	-0.43	0.14	0.002	-0.79	0.17	<0.001
Cigarette smoking, No	-0.58	0.24	0.013	-0.21	0.13	0.087	-0.67	0.17	<0.001	-1.02	0.21	<0.001
Tobacco smoking, No	0.33	0.22	0.138	-0.20	0.12	0.091	-0.07	0.18	0.685	-0.96	0.21	<0.001
Diabetes,	-8.69	0.17	<0.001	0.73	0.08	<0.001	7.54	0.20	<0.001	-0.48	0.16	0.002
Pre-diabetic												
Diabetic	-8.34	0.19	<0.001	16.84	0.22	<0.001	-6.19	0.13	<0.001	-0.33	0.19	0.086
Hypertriglyceridemia,	-0.73	0.14	<0.001	-0.03	0.08	0.738	-0.41	0.10	<0.001	0.23	0.13	0.071
Yes												
Hypercholesterolemia,	0.55	0.22	0.009	0.42	0.13	0.001	0.29	0.16	0.082	0.25	0.21	0.219
Yes												
BMI, normal	-0.73	0.35	0.043	-0.39	0.17	<0.001	-0.22	0.27	0.413	-0.92	0.31	0.003
Overweight	-1.58	0.37	<0.001	-0.58	0.18	0.001	-0.68	0.27	0.009	-1.48	0.31	<0.001
Obese	-1.89	0.36	<0.001	-0.76	0.18	<0.001	-1.01	0.27	0.002	-1.53	0.32	<0.001
Total Observation (n)		2144			4540			3127			9811	
R-squared		0.85			0.77			0.85			0.41	
MSE		8.50			4.95			6.47			31.04	

[DOI: 10.47176/mjiri.36.116]

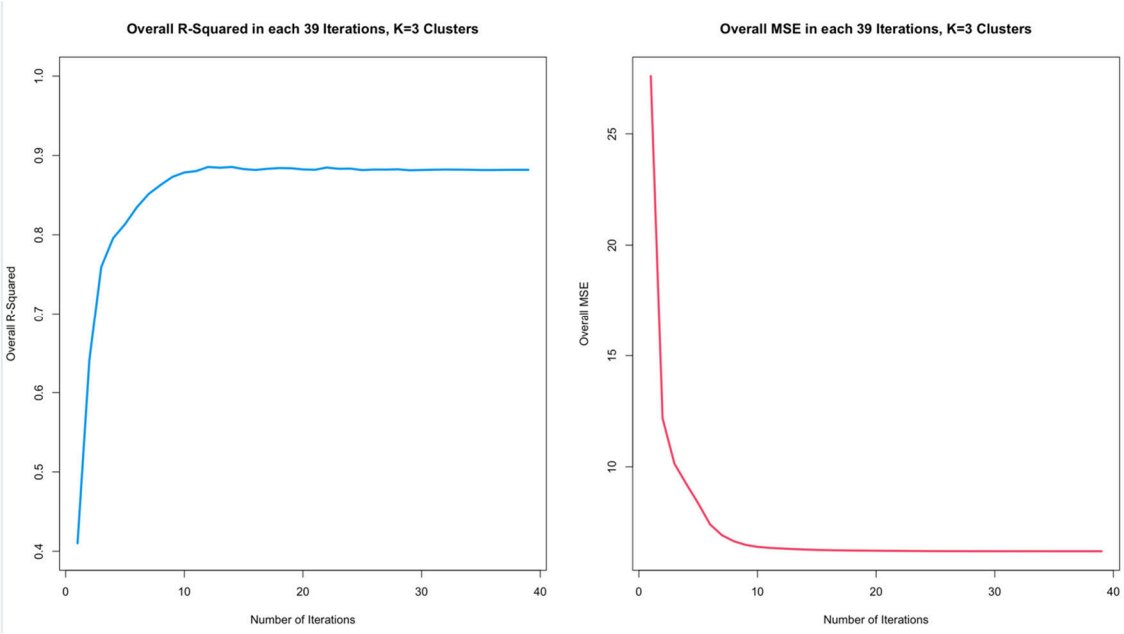

Fig. 1. Convergence curve for overall R-squared and overall MSE in 39 iterations

squared and the MSE values in all the 39 iterations of the CLR model. It could be clearly seen that the 39th iteration has the minimum overall MSE value.

Table 3 shows the ADS in different levels of predictor

Table 3. Average depression score of different levels of predictor variables in 3 clusters and all data

Variable	Cluster 1		Cluster 2		Cluster 3		All data	
	ADS± SD	p-value	ADS± SD	p-value	ADS± SD	p-value	ADS± SD	p-value
Sex		<0.001		<0.001		<0.001		<0.001
Male	11.5 ± 6.7		3.1 ± 3.5		8.9 ± 5.7		6.7 ± 6.1	
Female	15.1 ± 7.8		5.3 ± 5.2		12.1 ± 7.0		9.7 ± 7.7	
Marital Status								
Single	9.82 ± 4.7	<0.001	3.5 ± 3.3	<0.001	21.85 ± 7.5	<0.001	7.61 ± 7.4	<0.001
Married	15.17 ± 7.7		4.26 ± 4.7		9.94 ± 5.8		8.43 ± 7.1	
Divorced	14.71 ± 8.1		6.67 ± 4.5		27.78 ± 8.9		12.28 ± 9.6	
Widowed	8.95 ± 6.5		8.79 ± 6.0		17.53 ± 6.8		10.90 ± 7.3	
Physical Activity		0.023		<0.001		<0.001		<0.001
Low	13.76 ± 7.6		4.71 ± 5.0		11.30 ± 6.7		8.84 ± 7.3	
Medium	14.13 ± 7.7		4.39 ± 4.6		10.68 ± 6.7		8.57 ± 7.3	
High	12.73 ± 7.3		3.46 ± 3.7		9.70 ± 6.3		7.21 ± 6.7	
Cigarette Smoking		0.337		0.083		0.434		0.877
No	13.71 ± 7.7		4.43 ± 4.7		10.86 ± 6.7		8.49 ± 7.3	
Yes	14.27 ± 7.0		4.00 ± 4.4		10.55 ± 6.0		8.46 ± 7.0	
Tobacco Smoking		0.316		0.671		<0.001		0.001
No	13.70 ± 7.6		4.40 ± 4.7		10.68 ± 6.5		8.42 ± 7.2	
Yes	14.24 ± 7.6		4.29 ± 4.8		12.63 ± 7.7		9.27 ± 7.9	
Diabetes		<0.001		<0.001		<0.001		0.003
Normal	17.26 ± 7.8		3.69 ± 3.1		11.25 ± 6.0		8.52 ± 7.3	
Pre-Diabetic	9.90 ± 4.7		4.58 ± 4.0		19.9 ± 6.4		8.06 ± 6.7	
Diabetic	9.50 ± 5.0		23.41 ± 6.5		5.09 ± 4.7		9.00 ± 7.4	
Hypertriglyceridemia		<0.001		<0.001		0.001		0.012
No	14.43 ± 7.8		4.16 ± 4.2		11.14 ± 6.8		8.37 ± 7.3	
Yes	12.59 ± 7.2		5.01 ± 5.8		10.24 ± 6.3		8.76 ± 7.1	
Hypercholesterolemia		0.978		<0.001		0.504		0.002
No	13.76 ± 7.7		4.29 ± 4.5		10.86 ± 6.7		8.42 ± 7.2	
Yes	13.75 ± 7.2		5.55 ± 5.8		10.57 ± 6.5		9.24 ± 7.2	
BMI		0.002		<0.001		<0.001		<0.001
Underweight	15.28 ± 8.1		3.84 ± 2.9		14.30 ± 8.6		9.20 ± 8.3	
Normal	14.54 ± 7.7		4.03 ± 4.1		11.08 ± 6.8		8.36 ± 7.3	
Overweight	13.18 ± 7.4		4.33 ± 4.5		10.34 ± 6.3		8.20 ± 6.9	
Obese	13.60 ± 7.7		5.05 ± 5.7		10.85 ± 6.6		8.98 ±7.5	

[DOI: 10.47176/mjiri.36.116]

variables in the 3 clusters and all the data. In the marital status variable, the maximum ADS was 12.28 in the category of divorced people in all data. The maximum ADS for married persons was 15.17 in cluster 1, 8.79 for widowed people in cluster 2, and 27.78 for divorced people in cluster 3. In the diabetes variable, people with diabetic levels had the maximum ADS in all the data with the value of 9, yet the maximum ADS category differed in the 3 clusters. Healthy persons had the highest ADS with a value of 17.26, diabetic people had the highest ADS with a value of 23.41, and prediabetic people had the highest ADS with a value of 19.9. The underweight people had the highest overall ADS of 9.20, followed by the underweight people in clusters 1 and 3 with 15.28 and 14.30, respectively, for the BMI variable. Meanwhile, in cluster 2, obese people had the maximum ADS with the value of 5.05. ADS for the other variables are shown in Table 3.

Figure 2 demonstrates the ADS in the 3 clusters. The maximum ADS belonged to cluster 1 and the minimum ADS among the 3 clusters belonged to cluster 2. The

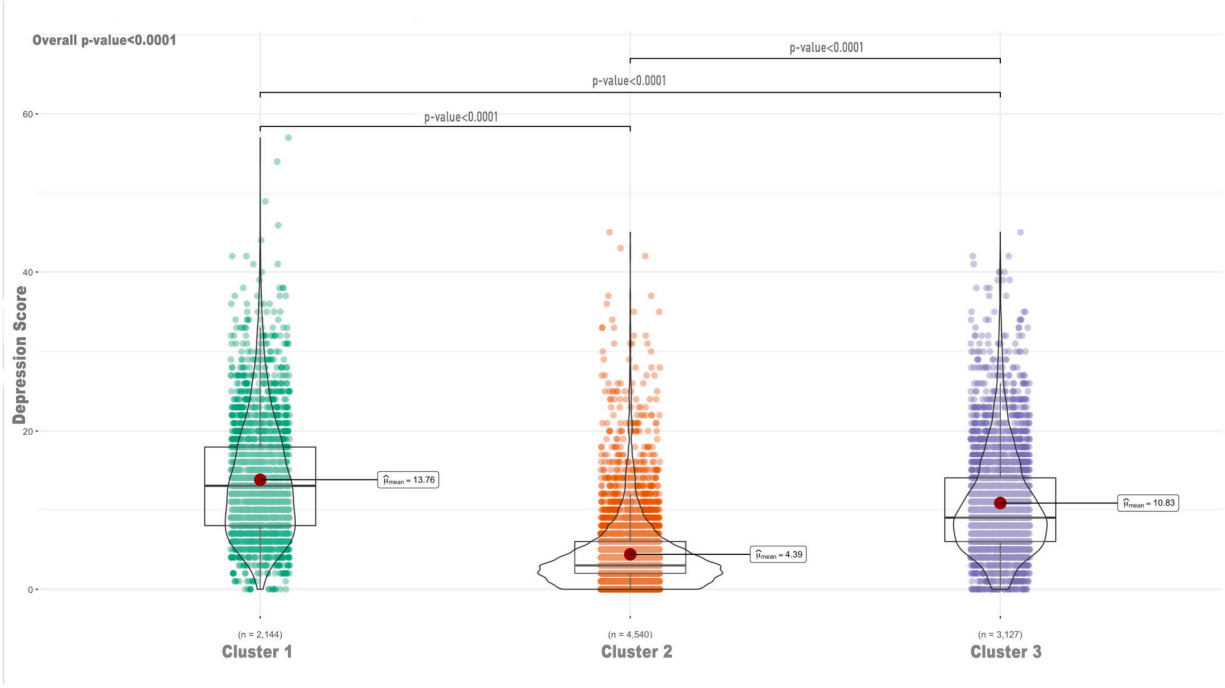

Fig. 2. Average of depression score by cluster

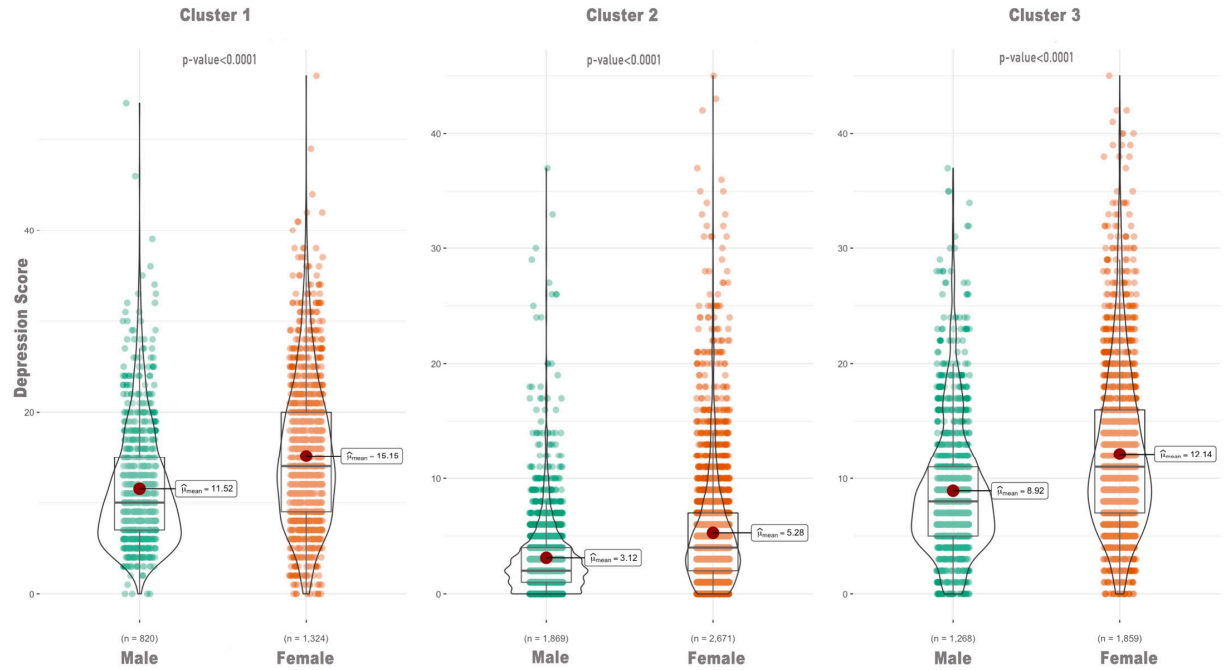

Fig. 3. Average of depression scores by sex, in three different clusters

[DOI: 10.47176/mjiri.36.116]

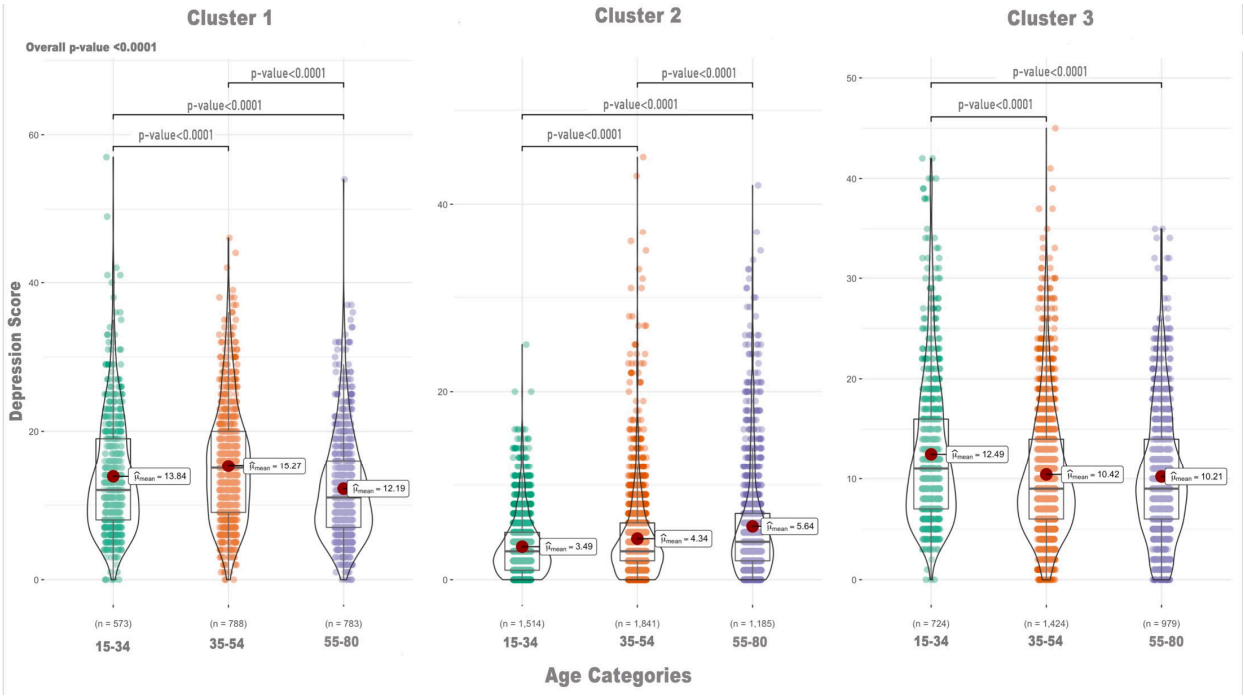

Fig. 4. Average of depression scores by age categories, in three different clusters

mean depression score was 13.76 ± 7.6 in cluster 1, 4.39 ± 4.7 in cluster 2, and 10.83 ± 6.7 in cluster 3. The overall equality test of ADS was significant with P = 0 and with pairwise comparison, we found statistically significant differences in ADS in any pairs of clusters (Fig. 2). Thus, to label our final clusters based on the findings of k = 3 clusters, it can be noted that the 3 clusters classify individuals into 3 degrees of ADS: low (cluster 2), moderate (cluster 3), and high (cluster 1). Figure 3 exhibits the ADS by sex in 3 clusters. In all the clusters, the ADS of women was significantly greater than that of men (P values are shown in Fig. 3). Finally, Figure 4 depicts the ADS by age categories in 3 clusters. In cluster 1, the age category of 35-54 years had the maximum ADS; in cluster 2, the age category of 55-80 years had the maximum ADS, and the maximum ADS belonged to the age category of 15-34 years in cluster 3. According to Figure 4, the overall equality of ADS in age categories was significantly different in all the clusters. Only the significant pairwise comparisons are shown in Figure 4.

Discussion

We conducted this cross-sectional analysis using data from the second phase of the Kerman Coronary Artery Diseases Risk Factors study (KERCADRS), which included 9811 individuals aged 15 to 80 years in 2014-2018.The relationship between depression and certain continuous and categorical predictor variables was investigated. Cluster-wise linear regression (CLR) was utilized to this end by clustering the individuals in the way that the multiple linear models in each cluster have the minimum MSE. The main benefit of the CLR method is that it helps to fit more accurate models. Particularly in big data with more heterogeneity, fitting only one model for the whole data set does not have enough accuracy and may lead to a high value of the MSE and a low value of the R-squared. In our study, we used both overall multiple linear regression and the CLR model and compared the results of these 2 methods. There were some variables with a nonsignificant effect on ADS in the overall model, but a significant effect on the CLR method and vice versa. Furthermore, the maximum ADS for the whole data was in the age category of 55-80 years; meanwhile, employing the CLR method, we found that in each cluster the age category with the maximum ADS was different.

Numerous researchers have examined depression data using a variety of clustering techniques. Tomita et al employed geographical clustering to examine food insecurity in South Africa and its connection with depression. They detected regional variability in food insecurity on a national scale in South Africa and revealed that hotspots had a higher incidence of incident depression (22). Based on the 17-item Hamilton Rating Scale for Depression baseline items, Kato et al used hierarchical cluster analysis with complete linkage to identify clusters of patients with major depressive disorder and to assess the efficacy of venlafaxine extended-release versus placebo, as well as the potential effect of dose on efficacy, in each cluster (23). Miller et al also used multilevel models to demonstrate clustering of depression and inflammation in teenagers who had previously faced adversity in childhood. Vicent-Gil et al used a 2-step clustering methodology to find homogenous patient categories. The first phase involves preclustering participants into tiny subgroups through a sequential clustering technique (24). The second phase takes the subclusters from the previous step as inputs and groups them into the optimal number of clusters using a hierarchical clustering method. Our method, CLR,

[DOI: 10.47176/mjiri.36.116]

has enough iterations to gain the minimum value of overall MSE and the individuals were clustered in the best way.

To discuss ADS in levels of other variables, we could mention the studies below, which we compared to our study. Panagiotakos et al showed that the average of Center for Epidemiologic Studies Depression Scale (CES-Depression) score is higher in never married people compared with married, divorced, and widowed people (25). St John et al also reported the relationship between marital status and depressive symptoms (26). In their study, the maximum CES-Depression score was observed in dissatisfied married people. In contrast, in our study, the maximum ADS belonged to divorced people in all the data and its level was different in the clusters as mentioned in the results section. Divorce has a detrimental impact on physical and psychological health on average for a variety of factors, including the social support received and the individual's financial status. Milic et al studied the relationship between cigarette smoking and depression of students in different locations of schools. The mean Beck Depression Index (BDI) score for ever-smoke students was statistically greater than never-smoke students (27). In our study, cigarette smoking had no significant effects on ADS neither in all the data nor in the clusters. Cigarette smoking and depression may have a bidirectional association, with occasional smoking initially reducing depression symptoms but eventually exacerbating them. Finally, there may in fact be no causal relationships between smoking and depression. Given the short half-life of nicotine, smokers may also report that cigarettes reduce their symptoms, which has led to the impression that smoking improves their mood. De Wit et al found a U-shape relationship between depression and 4 categories of BMI. It means that the depression score in underweight and obese people is more than the 2 middle categories (28). In our study, we also observed the same pattern in the overall data. However, after CLR method, in clusters 1 and 3, there was an almost decreasing pattern in ADS within BMI categories. It means that the maximum ADS was observed in underweight people and the minimum ADS was in obese people. In contrast, in cluster 2, an increasing trend was found, which means the obese people had the maximum ADS and underweight people had the minimum ADS. Being underweight is as harmful to the mental state as being fat. Thus, health researchers should be as attentive to underweight people since being underweight and obese both raise the risk of depression. Maintaining the appropriate weight through healthy eating and lifestyle is believed to be the best way.

The study's primary flaw is that it is a screening study that depends on standardized assessments rather than clinical evaluation by doctors utilizing diagnostic criteria. As a result, this is an anxiety and depression symptom survey. Of course, an epidemiological study with a high sample size would provide a platform for further research into the magnitude of the problem and the implementation of health intervention initiatives.

Conclusion

The findings of this cross-sectional study shed light on a number of characteristics linked to the average depression score. By grouping people into the appropriate groups, the CLR approach may identify the influence of these risk variables on ADS with more precision and accuracy. According to the findings, increasing public awareness of mental health issues and learning how to cope with them will change people's lives. In addition, it appears that doing periodic needs assessments is essential for identifying vulnerable populations. Figures 2 and 4 are particularly important because we can see that in the first cluster (Fig. 2), which has the highest level of depression score, there are 2144 people, which suggests that around 22% of people in the age group of 35-54 years have high levels of depression. The third cluster, which has a medium degree of depression (Fig. 2), includes 3127 people (32% of the total population) who are in the age bracket of 15-34 years and have the highest level of depression in this cluster. This can serve as a signal that people in the third cluster with moderate depression are frequently young people in the community, and we should consider measures to prevent depression in young people aged 15 to 34 years.

Acknowledgments

Not Applicable.

Abbreviations

ADS, Average Depression Score.
CLR, Cluster-wise Linear Regression.
MSE, Mean Squared Error.
BDI, Beck Depression Index.
BAI, Beck Anxiety Index.
GPAQ, Global Physical Activity Questionnaire.
FBS, Fasting Blood Sugar.
BMI, Body Mass Index.
SD, Standard Deviation.
KERCADRS, Kerman Coronary Artery Diseases Risk Factors Study.
MDD, Major Depressive Disorder.
VEN, Venlafaxine Extended-Release.
HAM-D17, HAMilton rating scale for Depression.
CES, Center for Epidemiologic studies Depression scale.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. Organization WH. Depression and other common mental disorders: global health estimates. World Health Organization; 2017.
2. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet. 2007;370(9590):841-50.
3. Christopher MS, Gilbert BD. Incremental validity of components of mindfulness in the prediction of satisfaction with life and depression. Curr Psychol. 2010;29(1):10-23.
4. Pine D. Integrating research on development and fear learning: a vision for clinical neuroscience? Depress Anxiety. 2009.

[DOI: 10.47176/mjiri.36.116]

5.Murray L, Hipwell A, Woolgar M, Cooper P. Cognitive vulnerability to depression in 5-year-old children of depressed mothers. J Child Psychol Psychiatry. 2001;42(7):891-9.
6. Alavi N, Aliakbarzadeh Z, Sharifi K, editors. Depression, anxiety, activities of daily living, and quality of life scores in patients undergoing renal replacement therapies. Transplant Proc.; 2009: Elsevier.
7. Akhondzadeh S, Tahmacebi-Pour N, Noorbala AA, Amini H, Fallah-Pour H, Jamshidi AH, et al. Crocus sativus L. in the treatment of mild to moderate depression: a double-blind, randomized and placebo-controlled trial. Phytother Res. 2005;19(2):148-51.
8. Mohammadi M-R, Davidian H, Noorbala AA, Malekafzali H, Naghavi HR, Pouretemad HR, et al. An epidemiological survey of psychiatric disorders in Iran. Clin Pract Epidemioly Ment Health. 2005;1(1):1-8.
9. Ghasemipour Y, Ghorbani N. Mindfulness and basic psychological needs among patients with coronary heart disease. Iran J Psychiatry Clin Psychol. 2010;16(2):154-62.
10. Rugulies R. Depression as a predictor for coronary heart disease: a review and meta-analysis. Am J Prev Med. 2002;23(1):51-61.
11. Najafipour H, Banivaheb G, Sabahi A, Naderi N, Nasirian M, Mirzazadeh A. Prevalence of anxiety and depression symptoms and their relationship with other coronary artery disease risk factors: A population-based study on 5900 residents in Southeast Iran. Asian J Psychiatr. 2016;20:55-60.
12. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007;52(1):258-71.
13. Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl Energy. 2015;160:153-63.
14. DeSarbo WS, Cron WL. A maximum likelihood methodology for clusterwise linear regression. J Classif. 1988;5(2):249-82.
15. Späth H. Algorithm 39 clusterwise linear regression. Computing. 1979;22(4):367-73.
16. Najafipour H, Mirzazadeh A, Haghdoost A, Shadkam M, Afshari M, Moazenzadeh M, et al. Coronary artery disease risk factors in an urban and peri-urban setting, Kerman, Southeastern Iran (KERCADR study): methodology and preliminary report. Ran J Public Health. 2012;41(9):86.
17. Beck AT, Steer RA, Brown G. Beck depression inventory–II. Psychol Assess. 1996.
18. Ghassemzadeh H, Mojtabai R, Karamghadiri N, Ebrahimkhani N. Psychometric properties of a Persian-language version of the Beck Depression Inventory-Second edition: BDI-II-PERSIAN. Depress Anxiety. 2005;21(4):185-92.
19. Steer RA, Beck AT. Beck Anxiety Inventory. In: C. P. Zalaquett, R. J. Wood, editors, Evaluating stress: A book of resources. Scarecrow Press. 1997;2:23–40
20. Hornik K. The comprehensive R archive network. Wiley Interdiscip Rev Comput Stat. 2012;4(4):394-8.
21. Patil I, Powell C. ggstatsplot:“ggplot2” based plots with statistical details. CRAN. 2018.
22. Tomita A, Cuadros DF, Mabhaudhi T, Sartorius B, Ncama BP, Dangour AD, et al. Spatial clustering of food insecurity and its association with depression: a geospatial analysis of nationally representative South African data, 2008–2015. Nature. 2020;10(1):1-11.
23. Kato M, Asami Y, Wajsbrot DB, Wang X, Boucher M, Prieto R, et al. Clustering patients by depression symptoms to predict venlafaxine ER antidepressant efficacy: Individual patient data analysis. J Psychiatr Res. 2020;129:160-7.
24. Miller GE, Cole SW. Clustering of depression and inflammation in adolescents previously exposed to childhood adversity. Biol Psychiatry. 2012;72(1):34-40.
25. Panagiotakos DB, Pitsavos C, Kogias Y, Mantas Y, Zombolos S, Antonoulas A, et al. Marital status, depressive episodes, and short-term prognosis of patients with acute coronary syndrome: Greek study of acute coronary syndrome (GREECS). Neuropsychiatr Dis Treat. 2008;4(2):425.
26. St John PD, Montgomery PR. Marital status, partner satisfaction, and depressive symptoms in older men and women. Can J Psychiatry. 2009;54(7):487-92.
27. Milic M, Gazibara T, Pekmezovic T, Kisic Tepavcevic D, Maric G, Popovic A, et al. Tobacco smoking and health-related quality of life among university students: Mediating effect of depression. PLoS One. 2020;15(1):e0227042.
28. De Wit LM, Van Straten A, Van Herten M, Penninx BW, Cuijpers P. Depression and body mass index, a u-shaped association. BMC Public Health. 2009;9(1):1-6.

[DOI: 10.47176/mjiri.36.116]