Analyzing Inter-relationship of Consumer Prices in UAE Using Multiple Linear Regression

Iman Saeed, Sarah Baras, and Jauhar Ali

***Abstract*—Today's nature of data requires a set of specialized techniques that can handle the huge and continuously growing amount of data. Regression analysis is one of the simplest techniques that predicts outputs and explains relationships in a given dataset. Regression also proves its ability to adapt to the new requirements of huge, multidimensional data. In this paper, we will use multivariable regression to study and predict the prices of a number of selected categories such as education and transportation. The data was collected by Dubai Statistical Center for monthly Consumer Price index in the Emirate of Dubai for the period from 2008 to mid-2016.**

***Index Terms*—Predicting, multiple linear regression, consumer price index, rapid miner.**

I. Introduction

In linear regression, there are two main motivations behind any statistical or data mining study. These motivations can be identified as understanding the dataset and predicting a future value. In the former, researchers look for some general statistics or theoretic explanations that can define the distribution of the values, the dimensionality of the dataset, and how much each attribute contributes to the overall correlation of the dataset. This most likely result in generating an explanatory model. In the later, a selected dataset will be analyzed to create some model to predict a future value [1]. Before creating the model, it is important to know that, the selected data must have at least two attributes. The first attribute provides a sense of what the model will be predicting, and it is known as the dependent variable, predicted, label or class name. The remaining attributes can be used as a base line for the model and are defined as predictors or independent variables. In case of a single predictor, a linear model can be mapped and explained by the following equation:

$$y = b_0 + b_1 x_1 \quad (1)$$

where (y) is the predicted value, (X) is the predictor and linearly related to (y), and b_0 is a constant that the corresponding value of "X" is "*0*", and b_1 is a regression weight that is used to determine the level of involvement of the predictor to the final model [2].

In general, data in real life are distributed in a way that makes it difficult to fit all the data to a linear line by using the previous equation. To improve the equation, the difference (ε) between the predicted value (y) and the actual value (y) was added to the equation. That difference is known as the residual, and it can be mathematically donated by $\varepsilon = (y - \bar{y})$ and the regression equation can be donated as $y = \bar{y} + (y - \bar{y})$ or simply as $y = b_0 + b_1 x_1 + \varepsilon$. Along with using one predictor, linear regression can also perform well with multivariable attributes. Equation (2) shows the way of extending linear regression to include more attributes:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...b_n x_n + \varepsilon \quad (2)$$

where n is the number of contributing predictors.

Correlation coefficient (R) is a related concept that we need to look at. Correlation is a statistical measurement that computes the relationships between attributes in the same dataset. The relationship may indicate a positive correlation which means an increase in an attribute reflects an increase in the positively correlated attribute. Negative is another relationship that reflects opposite correlation, so when one increases, the negatively correlated attribute will decrease. A zero correlation means there is no relationship between the attributes and their increase/ decrease has no effect on each other. The value of (R) is ranged between [-1, +1] and it is considered strong correlation when the value of (R) is close to ±1 [1, 2].The square value of R (Coefficient of determination) is used to check the quality of the predicted model. It helps to find the relationship between the variability of the predicted values and the actual values of the label. In other words, it is the percentage of data predicted that can be explained based on the predictors.

To highlight the most important contributors in linear regression equation, it is important to point that the weight value (b_n) in the equation expresses how much the value of (y) changes when (X_n) changes by one unit. It is also important to choose a weight value that minimizes the value of the residual (ε). In addition, residual errors in a predicting label values must be independent of each other and has no pattern or relation.

In this paper, we conducted an experiment by predicting prices of each Health and Education attributes in Consumer Price index (CPI) in emirate of Dubai by using values of other eleven attributes. We have used multiple linear regression algorithm to predict CPI in both attributes. We evaluated and analyzed the linear regression and performance results using two testing modes: Split data and Cross validation.

The rest of this paper is organized as follows. Section II is a literature review which is about contributions of other researchers in using multiple regression. In section III, we

Manuscript received March 26, 2018; revised May 15, 2018.

Iman Saeed, Sara Baras, and Jauhar Ali are with Abu Dhabi University, Abu Dhabi, UAE (e-mail: 1007458@students.adu.ac.ae, 1003299@students.adu.ac.ae, jauhar.ali@adu.ac.ae).

doi: 10.18178/joebm.2018.6.2.548

have preliminaries that include an explanation of the used data set and prediction tools and processes in our experiment. Section IV discusses the linear regression and performance results of the conducted experiment. We conclude the work in section V.

II. Litreture Review

Before moving on to discuss the approaches of linear regression, we need to point to the dimensionality of the dataset which is a critical area of study in data mining. A study by Cunningham and Ghahramani [3] discussed the topic and pointed to two areas of study: constrained (Orthogonal Matrix) and unconstrained objective. The orthogonal projection is mostly preferred over unconstrained Objectives because of their ability to visualize high dimensional data with a simple geometric interpretation. Fig. 1 summarizes the areas of dimensionality reduction described by them.

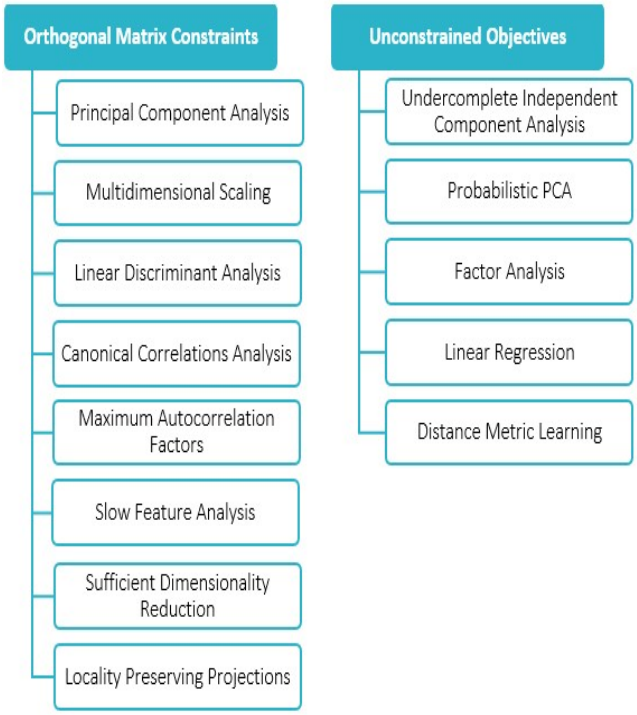

Fig. 1. Dimensionality reduction approaches.

In the multivariable regression, there are different approaches used to handle the dimensionality of dataset. Some researchers [4], [5] divided multivariable regression approaches into three approaches. The first one is the Simultaneous regression and recognized as predictive regression. The second approach is the Hierarchical regression or partitioning. The third approach is known as Stepwise regression. Each of these approaches has its own features and uses.

A. Simultaneous Regression

This type of regression treats all the attributes at the same time in one step. Because of that the estimated value is created all at once. Like the multivariable regression described before, the label will be predicted based on the predictors and the statistically estimated weights. In this method, all the predictors' contribution to the final model is evaluated from the beginning. The area where all the attributes are overlapping can cause a problem when creating a model as each predictor can claim its right to own it. This is usually handled by researchers by ignoring the whole overlapping area. That means that the ability of the model to predict a label can be affected as (R_2) in this case will equal to the values of the contributed part of each predictor excluding the overlapping part (the blue area). Fig. 2 illustrates the idea. The (Y) value is the predicted label, and both (X_1, X_2) are the predictors. The blue area defines (X_1, X_2) contributing area (R_2) while the gray area does not contribute to the final prediction model. The white area of (Y) is the residual or the data that cannot be explained by any of the predictors (1- R_2).

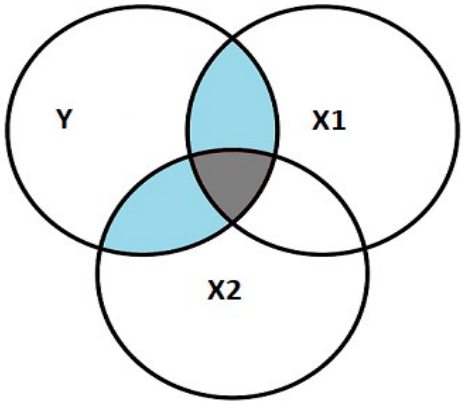

Fig. 2. Simultaneous regression.

Kohli and Gupta [6] used this approach by applying a set of different steps that can be aligned with this work schema. After selecting the targeted real time web data and preparing it, they executed multiple linear regression twice. They also performed another two iterations of sequential minimal optimization algorithm (SMO) which is a regression technique that aims to first normalize and then to standardize the data. The SMO algorithm was executed again but without applying any standardization or normalization for the dataset. The purpose of this research was to predict and to spot out any chance for growth. One may say this working flow may not be simultaneous, but our argument is that all attributes were given the same priority.

Another study followed the simultaneous approach to test linear regression ability to find patterns from muscular contraction or electromyogram (EMG). The study aimed to suggest simultaneous myoelectric control for different degrees of freedom (3-DoF) for the wrist/hand system. Hahne *et al.* [7] used linear and non-linear regression for the simultaneous and proportional predictors with 2-DoF for the wrist movements of the myoelectric control. They concluded that the performance of linear regression outperformed non-linear regression in terms of computational power when using good feature representation and regulation. They also concluded that nonlinear regression is more susceptible to over-fitting problem.

B. Hierarchical Regression

In this method researchers divide the processing of the predictors into steps. Each predictor set has different priority or execution order. In Fig. 3, we can deduce the execution

order of the predictors. We can ensure that (X_2) is executed before (X_1) and it has a higher contribution that affects the model as shown in the blue area. However, (X_1) is executed next and that is clear from the yellow area that belongs to (X_1). Also, we can say that none of the predictors' contribution is eliminated, and this makes (R_2) equals to the whole colored area (the blue and the yellow part together). Like in the simultaneous regression, the white area of (Y) cannot be explained by the predictors (X_1, X_2).

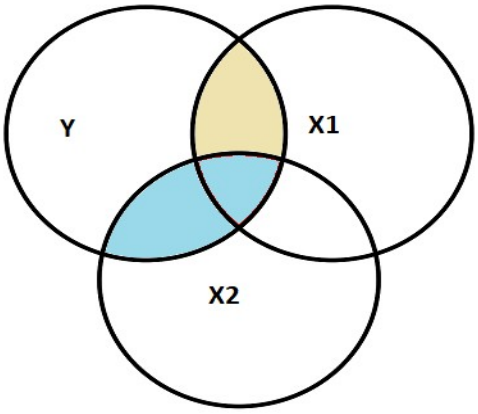

Fig. 3. Hierarchical regression.

Weng and Hwang [8] introduced an Incremental Hierarchical Discriminant Regression (IHDR) modeling that can withstand high dimensionality. Incremental Hierarchical learning algorithms requires the arrival of data in sequence at a given time. Along with the incremental knowledge gained, they aimed to automate the derivation of discriminating features from real-time data to build an incremental decision or regression tree. They believe that regression can outperform classification and this encouraged them to map classification to regression task which improves the performance of the created model. For testing the proposed algorithm, they ran three experiments with different datasets of high dimensionality that can reach to 5000 dimensions. They used sample in some cases that is less than the number of dimensions of the dataset. They separately tested datasets by using synthetic data, real face images, and data for autonomous navigation that inputs images and outputs steering signal.

Thukral *et al.* [9] also proposed a hierarchical regression approach. They used a dataset of face images. The used data was partitioned into different classes based on age. Then, a hierarchical regression executed the data groups to model a pattern. To improve the analyzed results, they used various types of classifiers and majority rule.

C. Stepwise Regression

Stepwise regression is another form of hierarchical regression though the difference lays under the method of attributes order selection. In this technique, researchers have no role in selecting the order of the predictors. The algorithm itself will set the order based on some statistical values such as the attribute or predictor with highest correlation with the predicted label. As in hierarchical regression, (R_2) will not exclude any attribute contribution to the model but will give the attribute tested first the priority [4], [5].

III. Preliminaries

A. Data Set

Our used data set is from Dubai Statistic Center, and it is for monthly Consumer Price index in the Emirate of Dubai from 2008 to mid-2016. It has 12 attributes with 103 instances of each of them. The attributes are as follows:

1) *Food and Non-Alcoholic Beverages Group*
2) *Alcoholic Beverages and Tobacco Group*
3) *Clothing and Footwear Group*
4) *Housing, Water, Electricity Gas and Other Fuels Group*
5) *Furnishings, Household Equipment and Routine Household Maintenance Group*
6) *Health Group*
7) *Transport Group*
8) *Communications Group*
9) *Recreation and Culture Group*
10) *Education Group*
11) *Restaurants and hotels Group*
12) *Miscellaneous Goods and Services Group*

Our objective is to predict the prices for each of Education and Health by using values of other attributes, and check which attribute contributes more in prediction and evaluate the performance in different data mining mechanisms.

B. Prediction Tool and Process

To do our experiment, we have used MS-Excel to prepare data and Rapid-Minor for data mining. We used multiple linear regression algorithm to predict consumer price index for each of Education and Health attributes. For evaluation the performance and the accuracy of our model, we did two different testing methods in predicting previous attributes by using same real dataset. The first one is holdout method, and the other one is cross validation.

Holdout method is also called Split Percentage because data set is divided in to ratios to conduct the evaluation. Split Percentage process is shown in Fig. 4, where 70% is training set that goes to data mining process for prediction modeling, and 30% is test set that is used in evaluation of the prediction model. Also, we used Split Data operator to manually perform a validation and divides data into test and training set based on percentages that we specified. In addition, we used linear regression operator in this process to calculate linear regression model from input data set that contains numerical values. We checked in eliminate collinear features and used none feature selection in regression and linear sampling in splitting data so data will be in consecutive order. Also, we checked the squared error, correlation, squared correlation and prediction average in performance operator.

Cross Validation method is also called k-fold because data set here is divided into k mutually exclusive subsets. Then, at i iteration a single subset is used as test set data while remaining is used as training set data. The value of k is usually 10 and we used this value. Fig. 5 shows the process in rapid minor, and we have used cross validation operator to estimate how accurately a model will perform in prediction. Both linear regression and performance operators are located inside the cross validation nested operator (Fig. 6), and we chose same criteria of previous method in both of them.

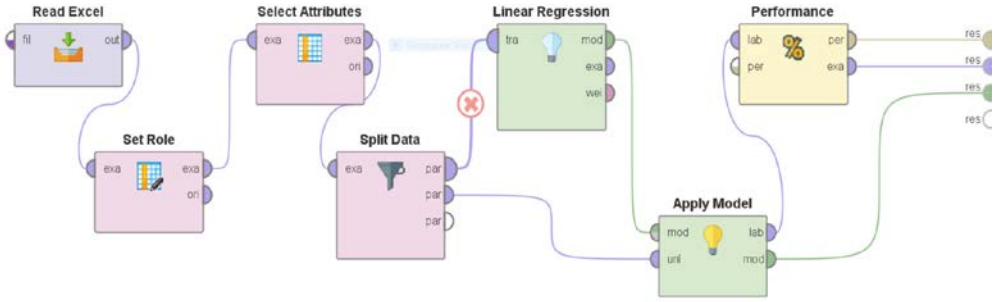

Fig. 4. Holdout (split) process.

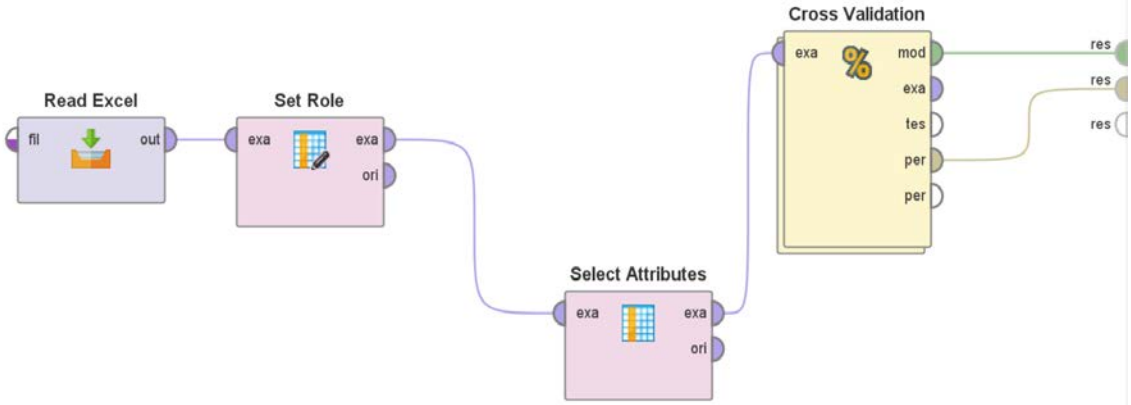

Fig. 5. Cross validation process.

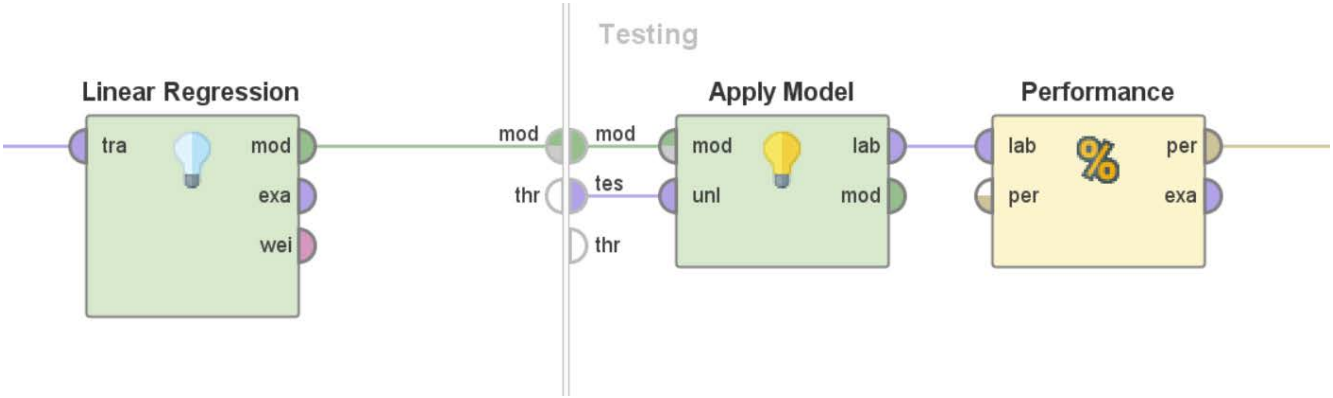

Fig. 6. Cross validation operator.

IV. Results

A. Linear Regression

We have conducted different experiments in evaluating the performance of selected tool in regression tasks. We started predicting consumer index prices of Health and Education by using two modes of testing.

We summarized the Linear Regression results of Health attribute in Table I. Rapid Miner uses asterisks under code label to indicate the significance of each attribute, so more stars implies to higher significance. The number of asterisks is actually based on *t*-Stat and *p*-Values. A low *p*-Value and high absolute value for *t*-Stat results in higher number of asterisks. In holdout (split) and Cross validation methods, we can see that restaurant & hotels and communication attributes have significant effect on health because their *p*-values are equal to 0 (less than 0.05). Also, they have the highest absolute *t*-Stat values compared to others. Therefore, any change in these attributes is related to significant change in health attribute. On the other hand, the Education attribute does not contribute much, so any change on it will have no or very little effect on the heath attribute.

In Table II, we present the results of linear regression of Education attribute. In both modes, Food and Non-Alcoholic Beverages, Clothing and Footwear, and Communications attributes are significant with health because their *p*-values are less than 0.05. As well, they have highest absolute *t*-Stat values. Consequently, any change in these three prices will have effect in health attribute price.

In addition, we have some attributes that have zero asterisks in code column in two modes like Housing, Furnishings, Health and transport. However, Alcoholic Beverages attribute has medium contribution in cross method while it appeared with zero contribution in split method. Besides, we have chosen the option of eliminating attributes that have correlated to each other in linear regression process in both methods, so attributes that correlated with each other

will be removed. Therefore, health and restaurants and hotels attributes are deleted in cross validation method because of previous reason while they are not removed in split because its results did not correlate.

TABLE I: HEALTH LINEAR REGRESSION

Attributes	t-Stat		p-Value		Code	
	Split	Cross	Split	Cross	Split	Cross
Food and …..etc	-0.249	-0.158	0.804	0.875		
Alcoholic Beverages …etc	2.276	2.866	0.026	0.005	**	***
Clothing ….etc	2.389	0.681	0.020	0.498	**	
Housing ,Water….etc	-0.105	-1.916	0.917	0.059		*
Furnishings ….etc	1.768	-1.942	0.082	0.055	*	*
Transport Group	1.766	2.263	0.082	0.026	*	**
Communications Group	-3.949	-4.495	0.000	0.000	****	****
Recreation and Culture Group	-2.426	-0.805	0.018	0.423	**	
Education Group	0.663	1.376	0.510	0.172		
Restaurants …etc	5.184	7.398	0.000	0.000	****	****
Miscellaneous …etc	1.833	1.800	0.072	0.075	*	*

TABLE II: EDUCATION LINEAR REGRESSION

Attributes	t-Stat		p-Value		Code	
	Split	Cross	Split	Cross	Split	Cross
Food and …..etc	4.728	4.696	0.000	0.000	****	****
Alcoholic Beverages …etc	2.163	1.207	0.034	0.231	**	
Clothing ….etc	3.366	3.152	0.001	0.002	***	***
Housing ,Water….etc	0.039	1.064	0.969	0.290		
Furnishings ….etc	0.071	0.045	0.944	0.964		
Health Group	0.663		0.510			
Transport Group	-1.454	-1.247	0.151	0.215		
Communications Group	-3.282	-3.939	0.002	0.000	***	****
Recreation and Culture Group	-0.774	3.010	0.442	0.003		***
Restaurants …etc	2.357		0.022		**	
Miscellaneous …etc	-1.789	2.149	0.079	0.034	*	**

B. Performance

```
performanceVector:
squared_error: 9.009+/- 7.210
correlation: 0.873
squared_correlation: 0.762
prediction_average: 125.087 +/- 0.453
```

Fig. 7. Health - performance of Holdout method.

```
performanceVector:
squared_error: 1.823+/- 1.208(mikro: 1.797 +/- 2.225)
correlation: 0.671 +/- 0.225 (mikro: 0.976)
squared_correlation: 0.500 +/- 0.248 (mikro: 0.953)
prediction_average: 118.877 +/- 5.979 (mikro: 118.880 +/- 6.079)
```

Fig. 8. Health - performance of cross method.

Rapid miner evaluates the performance of regression tasks by using different criteria. In prediction of health price attribute, when we used holdout method as in Fig. 7, it gave us good model with R_2 = 0.762. However, with cross validation as in Fig. 8, we had better model having average value of squared correlation that is closer to 1 (0.953).

Moreover, average squared error is less in cross validation method than in holdout method which indicates a better model too. Besides, the correlation of the model in cross validation mode indicates a perfect positive correlation with average value that is closer to 1 than in holdout mode. Prediction average of cross validation method is larger than holdout method prediction, and this is because prediction is done in different iterations (10 folds).

In the prediction of Education price attribute, it shows that squared correlation (= 0.886) that resulted in k-fold mode gave a better model compared with split mode that has value 0.369. However, squared error in holdout method, as in Fig. 9, is less than the squared error in cross validation, and we are looking for a model with less error.

Also, cross validation, as in Fig. 10, gave strong linear model with strong perfect positive correlation that is closer to 1 (= 0.941) while the correlation in holdout is weaker with value 0.607. The average of prediction in cross validation method is higher than holdout method prediction due to the different iterations in the first method.

```
performanceVector:
squared_error: 37.804 +/- 53.664
correlation: 0.607
squared_correlation: 0.369
prediction_average: 181.351 +/- 5.835
```

Fig. 9. Education - performance holdout method.

```
performanceVector:
squared_error: 52.665+/- 55.140(mikro: 52.351 +/- 91.647)
correlation: 0.470 +/- 0.258 (mikro: 0.941)
squared_correlation: 0.282 +/- 0.229 (mikro: 0.886)
prediction_average: 159.658 +/- 20.567 (mikro: 159.620+/- 20.878)
```

Fig. 10. Education - performance cross method.

V. CONCLUSION

This paper presented the results of experiment that is done on real monthly Consumer Price. We have used multivariable linear regression Algorithm to predict numerical values by using other attributes. Also, we used holdout and cross validation methods in testing, so we got different linear regression results with different performance. Furthermore, we have explained which result gave us better model in which case.

REFERENCES

[1] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining Concepts and Practice with RapidMiner*, New York: Elsevier, 2015

[2] R. W. Cooksey, "The methodology of social judgement theory," *Thinking & Reasoning Journal*, vol. 2, no. 2-3, pp. 141-174, 1996.

[3] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *Journal of Machine Learning Research*, vol. 16, pp. 2859-2900, 2015

[4] D. Wolff and M. L. Parsons, *Pattern Recognition Approach to Data Interpretation*, New York and London: Springer, 1983, p. 138.

[5] G. A. Morgan, J. A. Gliner, and R. J. Harmon, *Understanding and Evaluating Research in Applied and Clinical Settings*, New Jersey: Lawrence Erlbaum Associates, 2006.
[6] P. Thukral, K. Mitra, and R. Chellappa, "A hierarchical approach for human age estimation," in *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012, pp. 1529-1532.
[7] J. M. Hahne *et al.*, "Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 2, pp. 269-279, March 2014.
[8] J. Weng and W. S. Hwang, "Incremental hierarchical discriminant regression," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 397-415, March 2007.
[9] S. Kohli and A. Gupta, "Critical regression analysis of real time industrial web data set using data mining tool," *International Journal of Computer Applications* (*0975 – 8887), 4th International IT Summit Confluence 2013 - The Next Generation Information Technology Summit*, vol. 7, pp. 1-7, 2013.

Iman Saeed was born in Abu Dhabi, UAE. She received her B.Sc. in information system technology from Abu Dhabi University in Abu Dhabi in 2013, and she is currently a master student in information system technology in same university. She is currently an IT research assistant in Abu Dhabi University, and she worked in Banking sector for two years. Her research interests are ad-hoc network, data mining and mobile computing. She participated in UAE Graduate students Research Conference 2018.

Sarah Baras received a B. Sc. in computer science in 2012 from Abu Dhabi University, Abu Dhabi. She will be soon graduating from Abu Dhabi University with a master's degree in information technology. Currently, she is working as a research assistant in Abu Dhabi University, and she is contributing to the development of vehicular ad hoc network.

Jauhar Ali holds a master's degree in computer science from University of Peshawar, Pakistan. He received his PhD degree in computer science from University of Tsukuba, Japan in 1998. In the past, he served as assistant professor in UAE University, King Fahd University of Petroleum and Minerals, and Abu Dhabi University. He is currently serving as associate professor in computer science in Abu Dhabi University. His current research interests include data mining and visualization.