Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 2, June 2020, pp.991-999

Analyzing big data sets by using different panelized regression methods with application: surveys of multidimensional poverty in Iraq

Ahmed Mahdi Salih[1], Munaf Yousif Hmood[2]
[1]Dept. of Statistics, Wasit University
[2]Dept. of Statistics, University of Baghdad

ABSTRACT

Poverty phenomenon is very substantial topic that determines the future of societies and governments and the way that they deal with education, health and economy. Sometimes poverty takes multidimensional trends through education and health. The research aims at studying multidimensional poverty in Iraq by using panelized regression methods, to analyze Big Data sets from demographical surveys collected by the Central Statistical Organization in Iraq. We choose classical penalized regression method represented by the Ridge Regression. Moreover, we choose another penalized method, which is the Smooth Integration of Counting and Absolute Deviation (SICA) to analyze Big Data sets related to the different poverty forms in Iraq. Euclidian Distance (ED) was used to compare the two methods and the research conclude that the SICA method is better than Ridge estimator with Big Data conditions.

Keywords: Big Data, Penalized Regression, Poverty, Ridge Regression, SICA

Corresponding Author:

Ahmed Mahdi Salih
Dept. of Statistics Wasit University
Email: amahdi@uowasit.edu.iq

1. Introduction

Poverty in all its shapes is very important material to study due its effect over many psychological and economical properties of society, moreover; it is a very important variable in the demographic studies, demography science interested in population dynamics and the reason behind changing people composition and poverty [3]. Many researchers study poverty through demographical data because poverty itself is a multidimensional phenomenon related not only to financial conditions but also related to education and health conditions.

Therefore, demographical data provide searchers with vast information to study poverty. Demographical data are multi-types data that can be collected from many sources, there are three main sources for demographical data. The first is the digitized data from paper-based on demography over internet, the second is traces of social media since more than half the world population is social media sites users so it can make surveys that support demographical data, and the last one is the governmental offices like statistical departments and health agencies [4].

The variety in types and sources makes demographical data kind of Big Data as the following definitions "Extensive datasets, primarily in the characteristics of volume, velocity and/or variety, that require a scalable architecture for efficient storage, manipulation, and analysis" [16], "*Big Data is a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace*." [10]. Demographical data is big sets of data that require new and developed statistical methods to analyze. The study aims at using new techniques to deal with big data sets that can be transformed to be algorithms for application that gives fast and efficient results for big data sets analyzing. We are living in era that requires rapidity to take decisions and big data analysis should be fast and reliable system that can help managers to take big steps to develop establishments worldwide. Developing new and reliable system to deals with vast and huge amount of data is persistent need for many establishments in different sectors such as health, industry, marketing, etc. The study consists of nine sections, section 2 is related works, section

3 shows methods for Big Data analysis, section 4 introduce the Ridge regression, section 5 introduce the SICA method, section 6 shows the comparison method between estimators, section 7 introduces the demographical data under study and results, and section 8 is the conclusion of the study.

2. Literature background

Big Data issues and challenges attract researches around the world to introduce new statistical methods and techniques due to the fast development of technology and life at all aspects, Big Data have been studied by many researches such as.

Hoerl & Kennard [2] (1970) introduced a new regularized shrinkage estimator for the regression coefficients, in case of multicollinearity that appear in data with high dimensions and variety of data source, and they call it Ridge regression.

Tibshirani [12] (1996) suggested to use the L1-norm to develop a new penalty function to use on a regularized penalized optimization with specific conditions of high dimension and it is called LASSO estimator.

Knight & Fu [8] (2000) studied the asymptotic properties of the regular penalized estimator, and they presented an efficient estimator for the regression coefficients by developing penalty function works under Big Data conditions.

Lv & Fan [7] (2009) studied a family of penalty functions that depend upon the Lp-norm for the regression coefficients vector and they introduced new estimator with a mixed Lp-norms penalty functions for Big Data analysis.

Chudik at al [1] (2018) studied a sort of nonparametric estimators for the regression coefficients over a penalizing optimization and they introduced the OCMT One-Covariate at Time Multiple testing approach.

3. Methods for big data analysis

Knowledge about variables under study is the key issue to choose appropriate method of analysis either parametric or non-parametric method to analyze Big Data, many approaches have been submitted to analyze Big Data most of them aimed at reducing data dimension to avoid poor inference and bad performance of parameters under high dimensions conditions.

Reducing data dimensions attracts attention of many researchers over the world that they present different methods like penalizing over parameters or use appropriate prior distribution or select regressors to reduce high dimensions data into small sets of data to avert over fitting and improve forecasting. Collecting information and summarizing them in to a model is the first step in any statistical method parametric or non-parametric and selecting model depends on the nature of Big Data and the knowledge behind them [5].

Introducing numerous methods to summarize information from Big Data is the first step before taking an action in analysis like Principal Component Analysis, Factor Models, Sparse Principal Component Analysis and Partial Least Squares [8].

In our study, we select regression model with many explanatory variables in the form

$$\underline{y} = X\beta + \underline{\varepsilon} \quad \dots (1)$$

Where $\underline{y}$ represents (n × 1) independent variable vector and X is (n × p) explanatory variables matrix containing large number of variables and $\underline{\varepsilon}$ is (n × 1) random error vector and $\boldsymbol{\beta}$ is (p × 1) parameters vector. Regression models are commonly used in diverse statistical application with different kinds and types of data and sometimes researches choose regression models as starting models to improve them later or develop them in new kinds of models.

4. Ridge regression

Ridge regression is a kind of penalized regression which is simply a linear approach to deal with large sets of data, in equation (1) the basic idea of OLS method is estimate β that minimizes the errors ε where $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed random variables with mean equal to zero and variance σ^2. In other words, to find the estimators that minimize $\varepsilon'\varepsilon$ this optimization leads to the OLS estimators of parameters $\boldsymbol{\beta}^{OLS} = (X'X)^{-1}X'\underline{y}$ [24].

Under same assumption, the penalized regression minimizing errors subject to additional condition called penalty function.

$$\hat{\beta}^{PR} = \arg\min_{\beta} \frac{1}{n}(\varepsilon'\varepsilon + f(\lambda, \beta)) \quad \dots (2)$$

Where $f(\lambda, \beta)$ is the penalty function used to minimize the sum of squared errors where $0 < \lambda < 1$ in common searches $\lambda < 0.3$ is a complexity parameter that controls the amount of shrinkage the larger the value of λ, the greater the amount of shrinkage, there are many kinds of penalty functions that make a rise of different types of estimators. Ridge regression was first submitted by [2] Hoerl and Kennard 1970 by minimizing β throughout a Lp-norm penalty function $\|.\|_p$ where the Lp-norm is $\|\beta\|_p = \sum_{i=1}^{n} |\beta_i|^p$. Moreover, they used the L2-norm as a penalty function in the following form.

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \frac{1}{n}(\varepsilon'\varepsilon + \lambda I \|\beta\|_2) \quad .. (3)$$

Where $\boldsymbol{I}$ is $(p \times p)$ identity matrix and by solving the optimization in (3) we will apply shrinkage over β which minimize the sum of square errors, Ridge regression achieves sparse recovery and have some very good qualities and it is a good choice for high dimensions and Big Data analysis.
In terms of matrices, the optimization in (3) will be as follows.

$$L(\beta) = \left(\underline{y} - X\beta\right)' \left(\underline{y} - X\beta\right) + \lambda I \beta'\beta$$
$$L(\beta) = \underline{y}'\underline{y} - 2\beta X'\underline{y} + \beta' X' X\beta + \lambda I \beta'\beta$$

By differentiating with respect to β and equalize to zero [15].

$$0 = -2X'\underline{y} + 2X'X\beta + 2\lambda I\beta$$
$$X'\underline{y} = (X'X - \lambda I)\beta$$
$$\hat{\beta}^{Ridge} = (X'X - \lambda I)^{-1} X'\underline{y} \quad \dots (4)$$

Ridge regression is also highly recommended in case of multicollinearity problem.

5. Smooth integration of counting and absolute deviation

Smooth integration of counting and absolute deviation (SICA) method was first submitted by Lv &Fan 2009 [7] who presented a shrinkage method of penalized kind that meet model selection and sparse recovery problems. They started with the regression model in (1) and assumed that θ is the true regression coefficients vector, which theoretically goes to zero under shrinkage assumptions, but could be a small positive values to the true regression coefficients θ to be a nonzero vector that used to develop the penalty function. Penalized regressions usually use one order of Lp-norm for the parameter that the researchers wish to estimate like L2-norm in Ridge regression. Lv & Fan suggested to use mixture of Lp-norms that contain a ratio between L1 and L2-norm as a penalty function to make shrinkage over β, and they chose a penalty function that was studied by [9] Nikolova which is as follows.

$$f(\lambda, \beta) = \lambda \frac{(a+1)\|\beta\|_1}{a + \|\beta\|_1} \quad \dots (5)$$

Where $\boldsymbol{a}$ is a constant $\boldsymbol{a > 0}$ a very small number that is supposed to be non-negative small number [7]. It is clear from the penalty function in (5) that it is a ratio between two L1-norms and that makes the evaluation of β difficult somehow, there they suggested to upgrade the denominator in (5) to be quadratic and to use the true regression coefficients θ instead of β. The result is a penalty function that is equivalent to be a ratio between L1 and L2-norms and achieve the sparse recovery over β as follows.

$$f(\lambda, \beta) = \lambda \frac{(a+1)\|\beta\|_1}{(a + \|\theta\|_1)^2} \quad \dots (6)$$

Many approaches submitted to estimate θ some of them assume θ to be the $\hat{\beta}^{OLS}$, but the ordinary least square estimator assume $X'X$ to be full rank matrix with non-singular condition. Here, Lv & Fan suggested to use $\theta = \hat{\beta}^{OLS} = (X'X)^{+}X'y$ .Where + here represents the Moore-Penrose inverse or sometimes called Pseudoinverse which allows to solve any least squares system [11] for the matrices that have rank deficient by using minimum norm for each column in the solution matrix so if G is $(p \times p)$ matrix the Moore-Penrose inverse be as follows:

$$G^{+} = L(L'L)^{-1}(L'L)^{-1}L'G'$$

Where L is a simple extension of usual Cholesky factorization of non-singular matrices with removing zero rows, then L is a $(p \times r)$ matrix with rank equal to r where $G'G = LL'$. The **SICA** estimator is the solution of the following optimization [7].

$$\hat{\beta}^{SICA} = \arg\min_{\beta}(2^{-1}\varepsilon'\varepsilon + \lambda w\|\beta\|_1) \quad \ldots (7)$$

Where $w = \frac{(\boldsymbol{a+1})}{(\boldsymbol{a+\|\theta\|_1})^2}$ and by using the terms of matrices.

$$L(\beta) = \left(2^{-1}\left(\underline{y} - X\beta\right)'\left(\underline{y} - X\beta\right) + \lambda wc\beta'\right)$$

Where c is a $(p \times 1)$ ones vector and by simplifying the equation above we get:

$$L(\beta) = \left(2^{-1}(\underline{y}'\underline{y} - 2X'\beta + \beta'X'X\beta) + \lambda wc\beta'\right)$$

By differentiating with respect to β and equalize to zero we can evaluate :

$$\mathbf{0} = -X'\underline{y} + X'X\beta + \lambda wc$$
$$X'\underline{y} - \lambda wc = X'X\beta$$
$$\hat{\beta}^{SICA} = (X'X)^{+}\left(X'\underline{y} - \lambda wc\right) \quad \ldots (8)$$

6. Comparison method

Comparison among estimators is a main process in any statistical or scientific research that could help the researchers to determine the best statistical method of analyzing or model selection. Moreover, it helps them to create conclusions and take practical decisions. There are various statistical methods of comparison that are established over a specific assumption or theoretical groundings [14]. In statistics, the most common comparison method is the Mean Square Errors MSE but with high dimensions and the variety of data types and sources, MSE could lead to wrong imagination and poor understanding that could not help to create efficient decisions for the researchers. In our study, we have chosen the Euclidian Distance as a method of comparison due the high dimensions of data and the different types of data under study. Euclidian Distance is effective method to compare among estimators vectors and does not need any theoretical base to apply and its formula for a $(p \times 1)$ estimator vector as follows [6].

$$ED(\beta) = \frac{\sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2}}{p} \quad \ldots (9)$$

Euclidian Distance in (9) is divided by $\boldsymbol{p}$ that gives a weight to the victor dimension and as it gets smaller as better the estimator is.

7. Data and results

Before getting in to the data details we need to show short explanation to the concept of the Multidimensional Poverty Index MPI, this index was submitted by Alkire & Santos in 2011 [13] who studied various cases of

poverty and sorted them out into three main topics (Education, Health, and Living Standards). These topics are divided into 10 indicators and they are calculated for a group of families as follow.

Table 1. MPI calculation indicators

Indicators	**Households**				**Weights**
	1	**2**	**3**	**..**	
Household size					
Education					
1- No one has completed five years of schooling					**1/6**
2- At least one school-aged child not enrolled in school					**1/6**
Health					
3- At least one member is malnourished					**1/6**
4- One or more children have died					**1/6**
Living Standards					
5- No electricity					**1/18**
6- No access to clean drinking water					**1/18**
7- No access to adequate sanitation					**1/18**
8- House has dirt floor					**1/18**
9- Household has no car and owns at most one bicycle, motorcycle, radio, refrigerator, telephone or television					**1/18**
10- Household uses dirty cocking fuel					**1/18**
Score ci (sum of each deprivation multiplied by its weight					
Is the house hold poor (c > 0.33)					
Censored data c(k)					

MPI calculation depends upon two main variables, first one is H which is called the proportion of incidence, which is as following.

$$H = \frac{q}{n} \quad \dots (10)$$

Where q is the number of the people who suffer from multidimensional poverty, and n represents the total number of people in the group of families.
Second variable is A, it is called the intensity of poverty which is the average of deprivation of multidimensional poverty and can be expressed as:

$$A = \frac{\sum_{i=1}^{n} c_i(k)}{q} \quad \dots (11)$$

Where finally the Multidimensional Poverty Index is.

$$MPI = HA \quad \dots (12)$$

We have gotten sets of surveys data from the Central Statistical Organization IRAQ represent 300 group of families, and from it we get the MPI vector (300×1) by equations (10),(11),(12) and it here represents our independent variable vector y. We also have gotten data from surveys for the same groups for age and sex and many biological and social properties, we have 100 variables from different types quantitative, ordinal, nominal …etc., as detailed and classified in Appendix. Then these variables

will represent the explanatory variables matrix X with (300 × 100) where $n = 300, p = 100$. Estimators is evaluated for both Ridge and SICA methods according to equations (4),(8) and ED was calculated according to (9). We assume $a = 1 \times 10^{-4}$ [7] and we choose two values for λ it is respectively $[0.15, 0.3]$.

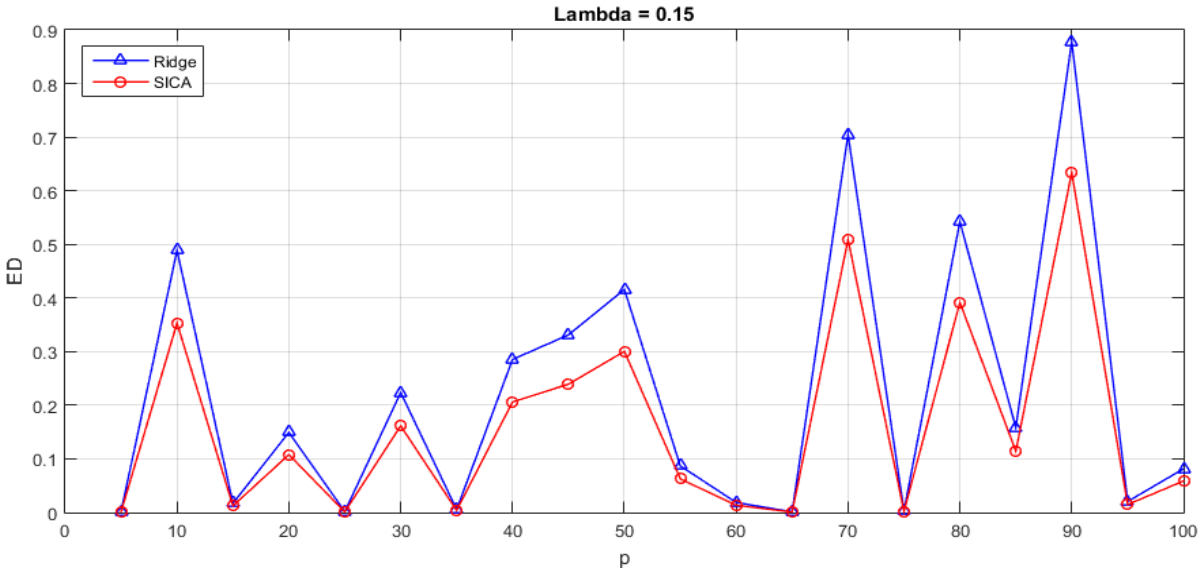

Figure 1. Euclidian Distance ED for Ridge and SICA methods, $\lambda = 0.15$

From Figure 1, when $\lambda = 0.15$ and the values of ED for the both methods, we can notice that the Ridge Estimator is very close to the SICA estimator in many points of p (number of variables) , but when $p > 75$ we can see that SICA estimator gives a better performance than Ridge estimator. In addition, the difference between the two methods get quite bigger as p goes larger.

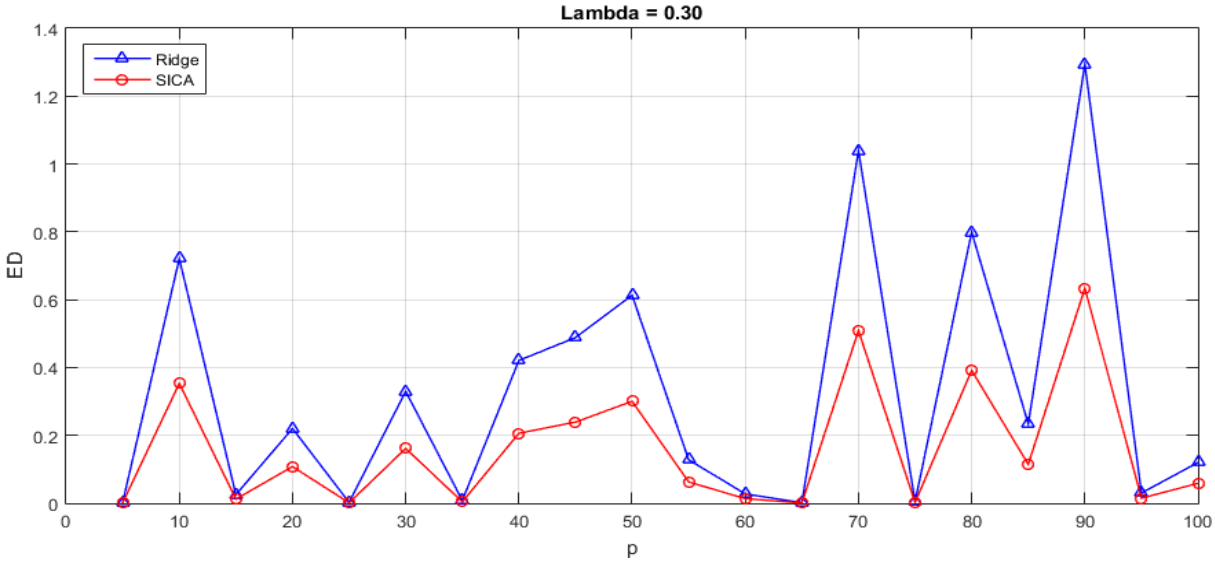

Figure 2. Euclidian distance for Ridge and SICA methods, $\lambda = 0.3$

From Figure 2, when $\lambda = 0.3$ we can see that the Euclidian Distance ED get quite bigger between the two methods Ridge estimator and SICA estimator when p gets bigger as we add more variable to the sample.

Table 2. The Euclidian distance values for both methods

	$\lambda = 0.15$		$\lambda = 0.3$			$\lambda = 0.15$		$\lambda = 0.3$	
p	**Ridge**	**SICA**	**Ridge**	**SICA**	**p**	**Ridge**	**SICA**	**Ridge**	**SICA**
5	**0.0020**	**0.0014**	**0.0029**	**0.0014**	**50**	**0.4159**	**0.3005**	**0.6132**	**0.3004**
10	**0.4888**	**0.3532**	**0.7208**	**0.3531**	**55**	**0.0869**	**0.0628**	**0.1281**	**0.0628**

	$\lambda = 0.15$		$\lambda = 0.3$			$\lambda = 0.15$		$\lambda = 0.3$	
15	**0.0171**	**0.0124**	**0.0252**	**0.0124**	**60**	**0.0187**	**0.0135**	**0.0275**	**0.0135**
20	**0.1494**	**0.1079**	**0.2203**	**0.1079**	**65**	**0.0012**	**0.0008**	**0.0017**	**0.0008**
25	**0.0016**	**0.0012**	**0.0024**	**0.0012**	**70**	**0.7043**	**0.5088**	**1.0385**	**0.5088**
30	**0.2237**	**0.1616**	**0.3298**	**0.1616**	**75**	**0.0031**	**0.0023**	**0.0046**	**0.0023**
35	**0.0055**	**0.0040**	**0.0081**	**0.0040**	**80**	**0.5421**	**0.3916**	**0.7993**	**0.3916**
40	**0.2854**	**0.2062**	**0.4208**	**0.2062**	**85**	**0.1584**	**0.1145**	**0.2336**	**0.1145**
45	**0.3313**	**0.2393**	**0.4884**	**0.2393**	**90**	**0.8773**	**0.6338**	**1.2935**	**0.6338**
50	**0.4159**	**0.3005**	**0.6132**	**0.3004**	**100**	**0.0815**	**0.0589**	**0.1202**	**0.0589**

8. Conclusions

Based on the employed data and results, we can determine that the both methods are equivalent somehow when the number of variables p are quite small relative to the sample size, and when p gets large relative to the sample size we can determine that SICA estimator is notably better than Ridge estimator with Big Data conditions.

References

[1] A. Chudik, G. Kapetanios and M. Pesaran, "One-Covariate at Time, Multiple Testing Approach to variable selection in High-Dimensional Regression Models", *Econometrica*, Vol. 86, Issue. 4, pp. 1479-1512, 2018.

[2] A. Hoerl and R. Kennard,"Ridge Regression: Biased Estimation for Nonorthogonal Problems" , *Technometrics*, Vol. 12, No. 1, pp. 55-67, 1970.

[3] D. Acharjya and A. Kauser, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", *International Journal of Advanced Computer Science and Applications*, Vol. 7, No 2, pp. 511-518, 2016.

[4] D. Alburez-Gutireerez , S. Aref, S. Gil-Clavel, A. Grow, D. Negraia and A. Zagheni, "Demography in the Digital Era: New Data Sources for Population Research", Smart statistics for smart applications: Book of short papers, pp. 23-30, Pearson Inc., ITALY, 2019.

[5] G. Kapetanios, M. Marcellino and K. Petrova, "Analysis of the Most Recent Modeling Techniques for Big Data with Particular Attention to Bayesian Ones", Eurostat. Statistical working papers. ISBN 978-92-79-77350-1, 2018.

[6] J. Fan and H. Fang, "Challenges of Big Data Analysis", *National Science Review*, No.1, pp. 293-314, 2014.

[7] J. Lv and Y. Fan, "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares" *The Annals of Statistics*, Vol. 37, No. 6A, pp. 3498-3528, 2009.

[8] K. Knight and W. Fu, "Asymptotics for Lasso-Type Estimator", *The Analysis of Statistics*, Vol. 28, No. 5, pp. 1356-1378, 2000.

[9] M. Nikolova, "Local Strong Homogeneity of Regularized Estimator", *SIAM Journal of Applied Mathematics*, Vol. 61, No. 1, pp. 633-658, 2000.

[10] M. Schroeck , R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "*Analytics: The Real-World Use of Big Data*", IBM Global Services Route 100 Somers, NY 10589 U.S.A, pp. 1-19, 2012

[11] P. Courrieu, "Fast Computation for Moore-Penrose Inverse Matrices", *Neural Information Processes – Letters Reviews*, Vol. 8, No. 2, pp. 25-29, 2005.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso*", Journal of Royal Statistics Society. B*, Vol. 58, No. 1, pp. 1456-1490, 1996.

[13] S. Alkire and M. Santos, "The Multidimensional poverty Index (MPI)" MPI: Construction and Analysis, ophi@qeh.ox.ac.uk, 2011.

[14] S. Mishra, V. Dhote, G. Prajapati and J. Shukla, "Challenges in Big Data Application: A Review", *International Journal of Computer Applications*, Vol. 121, No 19, pp. 42-46, 2015.

[15] T. Hastie, R. Tibshirani and J. Friedman, "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*", *Springer Series in Statistics*, USA, 2010.

[16] W. Chang and N. Grady, "*NIST Big Data Interoperability Framework: Volume 1, Definitions*" Special Publication (NIST SP) - 1500-1 Version 2, 2012.

Appendix: Demographical Data details.
N: Nominal - Q: Quantitative – O: Ordered – B: Binary (0,1)

N	Variable	Type	N	Variable	Type
1	Household type	N	26	Highest grade completed at that level	B
2	Line number	Q	27	Age 4-24	Q
3	Relationship to the head	Q	28	Check: Ever attended school or any Early Childhood Education programmed	B
4	Sex	N	29	Attended school during current school year(2017-2018)	B
5	Month of birth	O	30	Level of education attended current school year(2017-2018)	N
6	Age	Q	31	Grade of education attended current school year(2017-2018)	O
7	Line number of woman age 15 - 49	Q	32	Attended public school current school year(2017-2018)	B
8	Line number of man age 15 - 49	Q	33	School tuition in the current school year	B
9	Line number for children age 0-4	Q	34	Material support in the current school year	B
10	Member age 0-17	Q	35	Attended school previous school year(2016-2017)	B
11	Is natural mother alive	B	36	Level of education attended previous school year(2016-2017)	B
12	Does natural mother live in HH	B	37	Grade of education attended previous school year(2016-2017)	O
13	Natural mother's line number in HH	Q	38	Day of interview	Q
14	Where does natural mother live	B	39	Month of interview	O
15	Is natural father alive	B	40	Area	N
16	Does natural father live in HH	B	41	Region/Governorate	N
17	Natural father's line number in HH	Q	42	Region	N
18	Where does natural father live	B	43	Mother's line number	Q
19	Line number of mother or primary caretaker for children 0-17 years of age	Q	44	Father's line number	Q
20	Line number	Q	45	Education of household head	O
21	Age	Q	46	Functional difficulties	B
22	Age 4 and above	B	47	Health insurance	B
23	Ever attended school or Early Childhood Education programmed	B	48	Age at beginning of school year	Q
24	Highest level of education attended	N	49	Mother’s education	O
25	Highest grade attended at that level	N	50	Mother's functional disabilities (age 18-49 years)	B

Appendix: Demographical Data details (continued)

N	Variable	Type	N	Variable	Type
51	Mother's functional disabilities (age 18-49 years)	B	76	child no attending	B

52	Father’s education	O	77	any child not attend	B
53	Household sample weight	Q	78	Household has all school age children up to class 8 in school	B
54	Combined wealth score	R	79	Woman's line number	Q
55	Wealth Quintile	O	80	Women BH	B
56	Percentile Group of com1	Q	81	Total child death for each women (birth recode)	Q
57	Urban wealth score	Q	82	Total child death for each women in the last 5 years (birth recode)	Q
58	Wealt Quintile Urban	Q	83	Result of woman's interview	O
59	Percentile Group of urb1	Q	84	Ever given birth	B
60	Rural wealth score	Q	85	Ever had child who later died	B
61	Wealth Quintile Rural	Q	86	Boys dead	B
62	Percentile Group of rur1	Q	87	Girls dead	B
63	Primary sampling unit	Q	88	Women WM	B
64	Stratum	Q	89	Martial	B
65	Household ID	Q	90	Native language of the Respondent	B
66	Individual ID	Q	91	Translator used	B
67	Highest educational level attended	O	92	Rank number of the selected child	O
68	Highest year of education completed	O	93	Child line number	O
69	Total number of years of education accomplished	Q	94	Child's age	Q
70	Child education u 6	Q	95	Consent for interview girls 15-17	Q
71	years of education u 6	Q	96	Consent for interview boys 15-17	Q
72	Household has at least one member with 6 years of education	Q	97	Consent for Water Quality Testing	B
73	Attended school during current school year	B	98	Respondent to HH questionnaire	B
74	child schooled	B	99	Number of HH members	Q
75	No missing school attendance for at least 2/3 of the school aged children	B	100	Native language of the Respondent	B