Int. J. Nonlinear Anal. Appl. 12 (2021) No. 2, 931-938
ISSN: 2008-6822 (electronic)
http://www.ijnaa.semnan.ac.ir

Big Data Analysis by Using One Covariate at a Time Multiple Testing (OCMT) Method: Early School Dropout in Iraq

Ahmed Mahdi Salih[a,*], Munaf Yousif Hmood[b]

[a]*Department of Statistics, College of Administration and Economics Administration and Economics, Wasit University, Wasit, Iraq*
[b]*Department of Statistics, College of Administration and Economics Administration and Economics, University of Baghdad, Baghdad, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

The early school dropout is very significant portents that controls the future of societies and determine the nature of its elements. Therefore, studying this phenomenon and find explanations of it is a necessary matter, by finding or developing appropriate models to predict it in the future. The variables that affect the early school dropout Iraq takes a large size and multiple sources and types due the political and economic situation, which attributes it as a sort of Big Data that must be explored by using new statistical approaches. The research aims at using one Covariate at a Time Multiple Testing OCMT Method to analyze the data from surveys collected by the Central Statistical Organization IRAQ, which contains many indicators related to school dropout and meaningfully affect the life of the Iraqi persons. The Ridge Regression Method as well as the OCMT method were chosen to analyze data and the Mean Square Errors MSE was used to compare the two methods and From the results we find that OCMT estimator is better than Ridge estimator with Big Data conditions.

Keywords: Big Data, OCMT, ridge regression, multiple testing.

1. Introduction

Early school dropout is a foremost topic to study and analyze due its effect over a numerous properties of society political, psychological and financial, moreover; its represents an important

*Corresponding author
Email addresses: amahdi@uowasit.edu.iq (Ahmed Mahdi Salih), munaf.yousif@coadec.uobaghdad.edu.iq (Munaf Yousif Hmood)

Received: April 2021 *Revised:* May 2021

variable in the demographic studies, demography science interested in population dynamics and the reason behind changing people composition and poverty [1]. Early school dropout bring the courtesy of many researchers over the word because lack of education is a main reason of poverty and ignorance inside the society. Demography surveys in many of its shapes deals with the educational levels and school attending.

Therefore, demographical data afford searchers with massive information to analyze school dropout. Demographical data can be calculated from many sources like health agencies of official organizations [6] and this kind of data is complex and containing many kinds of data like ordinal, binary, quantitative, etc ...

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 [2]. Demographical data is big sets of data that need new statistical methods to study and analyze. The study aims at using new statistical methods to analyze Big Data in demography field. The study consists of seven sections, section 2 is Big Data Analysis, section 3 shows One Covariate at a Time Multiple Testing OCMT, section 4 introduce the Ridge Regression, section 5 introduce Mean Squares Error MSE, section 6 introduces data under study and results, and section 7 is the conclusion of the study. Big Data issues and challenges attract researches around the world to present new statistical methods and procedures due to the fast development of technology and life at all aspects, Big Data have been studied by many researches such as.Hoerl & Kennard [7] (1970) presented a new shrinkage estimator, in case of multicollinearity that seem in data with high dimensions and variety of data source, and they call it Ridge regression. Tibshirani [10] (1996) recommended to use the L1-norm to grow a new penalty function to use with specific conditions of high dimension and it is called LASSO estimator. Lv & Fan [10] (2009) studied a group of penalty functions that be contingent upon the Lp-norm and they announced new estimator with a mixed Lp-norms penalty functions for Big Data analysis. Chudik at al [3] (2018) studied a kind of nonparametric estimators for the regression coefficients over a penalizing optimization and they presented the OCMT One-Covariate at Time Multiple testing method.

2. Big Data Analysis

Information about variables under study is the main concern 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.

Decreasing data dimensions entices attention of many researchers over the world that they create 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 avoid over fitting and improve forecasting.

Introducing many methods to abridge 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 [9].

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

$$\underline{y} = X\beta + \underline{\epsilon}, \tag{2.1}$$

where $\underline{y}$ denotes (n $\times$ 1) independent variable vector and X is (n $\times$ p) explanatory variables matrix containing large number of variables and $\underline{\epsilon}$ is (n $\times$ 1) random error vector and $\beta$is (p $\times$ 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 recover them later or develop them in new kinds of models.

3. One Covariate at a Time Multiple Testing OCMT

Many approaches of Big Data analysis focus on penalizing regression, which take the largest place in Big Data analysis, as an alternative many researchers developed methods that focus on combining penalizing regression with greedy algorithms technique. In this method the authors focus on the predictive power of individual regressors instead of all the variables in the sample, which they call it Greedy Methods. In such methods, regressors chosen one by one based on their ability to represent the dependent variable.

OCMT method were submitted in 2016 by Chudik, Kapetanios and Pesaran [8] they focus on the overall impact of the x_i rather than the marginal impact $I(\beta \neq 0)$ they give attention to the overall impact of regressors and combine them with the marginal impact of regressors and they depend on multiple testing overall impact of regressors can be given by.

$$\theta = \sum_{i=1}^{n} I\beta_i\sigma_{ij} \qquad \beta \neq 0 \tag{3.1}$$

where β_i and σ_{ij} are OLS estimators for the model parameters, we have four possibilities for the impact.

	$\theta \neq 0$	$\theta = 0$
$\beta \neq 0$	***Case 1***	***Case 2***
$\beta = 0$	***Case 3***	***Case 4***

And by focusing on the overall impact it's clear that case (1,2,3) represent the overall and the marginal impact of regressors, the problem is that how we can select regressors that attend marginal and that overall impact.

If we suppose k represents the number of variables with $(\beta \neq 0)$ and k^* represent the other remain variables $(p-k)$ then the procedure of OCMT estimator based on the following assumption [3].

Assumption

Let $X_{k,k^*} = (X_k, X^*_{k^*})$ where $X_k = (x_1, x_2, \ldots, x_k)$ and $X^*_{k^*} = (x_{k+1}, x_{k+2}, \ldots, x_{k+k^*})$ are $n \times k$ and $n \times k^*$ observations matrices on signals and noise variables, and suppose that there exist n_0 such that for all $n > n_o, (n^{-1}X'_{k,k^*}X_{k,k^*})^{-1}$ is nonsingular with its smallest eigenvalue uniformly bounded away from zero. Therefore, there is an iterative selection method for regressors by using greedy algorithm submitted to be easy, fast, and efficient for large data sets. It's begin with choosing k regressors then add to them until we have the regressors that have the overall impact on Y.

The OCMT procedure depends on testing each estimated parameter for each variable in the model 2.1 by applying student t test for each of regressors [8].

$$t_{\beta_i} = \frac{\beta_i}{se(\beta_i)}, \tag{3.2}$$

where $\beta = (X'X)^{-1}X'Y$ is simply the OLS estimator for, and $se(\beta_i)$is the standard error for β_i.

Then we choose the regressors with $|t_{\beta_i}| > C_{p1}$ where represent the critical value for student-t test.

Let we assume that we choose k regressors in the first stage. The critical value size C_{p1} is very important for the second stage and with modification of type 1 error we can establish a new critical value for the next stage as follows [3].

$$C_{p2} = \exp\left[-\frac{(C_{p1})^2}{2}\right]. \tag{3.3}$$

Then we repeat the student-t test for each of the remaining (p-k) regressors as follows

$$t_{\beta_j} = \frac{\hat{\beta}_j}{se\ (\hat{\beta}_j)}, \qquad j = k+1, k+2, \ldots, p. \tag{3.4}$$

Then we choose the regressors with $|t_{\beta_i}| > C_{p2}$ and add them to the k regressors were chosen in the first stage, and we repeat this procedure by changing the critical value size until we have no regressors to add.

We can write the steps into Algorithm as follows

1. Test each of regressors with student-t test $t_{\beta_i} = \frac{\beta_i}{se(\beta_i)}$.
2. Choose the k regressors that attend $|t_{\beta_i}| > C_{p1}$.
3. Change the critical value size $C_{p1} = exp[-\frac{(C_{p1})^2}{2}]$.
4. Test the remaining (p-k) regressors with student-t test $t_{\beta_j} = \frac{\beta_j}{se(\beta_j)}$.
5. Choose the regressors that attend $|t_{\beta_i}| > C_{p2}$.
6. Add the new chosen regressors to the k regressors in step 2.
7. Repeat the procedure until we have no regressors to add.

In the end, we will get the effective variables that have marginal and overall impact

$$\beta^{OLS} = \begin{cases} \beta^{OLS} & if I(\beta \neq 0) = 1 \\ 0 & otherwise \end{cases} \tag{3.5}$$

Where $\beta^{OLS}$is the ordinary least square estimator.

4. Ridge Regression

Estimating the parameters of the model in (2.1) require an efficient estimation method, that can endures all the problems in data. Classical methods such as Least Squares and Maximum Likelihood methods oblige some assumptions in the model (2.1) such as $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$ that is hard to attend in real data. Moreover; the correlation matrix of X must be near unit matrix. And these assumption will be hard to attend in Big Data sets due the variety of the data types and different sources of it. In case of not attending these assumptions, the Least Square and Maximum Methods will lead to inefficient estimates.

Ridge regression is from early methods that were recommended to analyze large sets of data it is a kind of panelized regression which is simply a linear approach to deal with large sets of data, in equation (2.1) the basic idea of OLS method is estimate β that minimizes the errors ϵ where $\epsilon_1, \ldots, \epsilon_n$ are independent identically distributed random variables with mean equal to zero and variance σ^2. In other words to find the estimators that minimize $\epsilon^{'}\epsilon$ this optimization leads to the OLS estimators of parameters $\beta^{OLS} = (X'X)^{-1}X'Y$ [7].

Under the conditions of large data sets, the penalized regression minimizing errors subject to additional condition called penalty function.

$$\beta^{PR} = arg\min_{\beta} \frac{1}{n}(\epsilon^{'}\epsilon + f(\lambda, \beta)) \tag{4.1}$$

Where $f(\lambda, \beta)$ is the penalty function that used to minimizes the sum of squared errors where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage the larger the value of λ, the greater the amount of shrinkage, there is many kinds of penalty functions that make a rise of different types of estimators. Ridge regression first submitted by [7]. 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.

$$\beta^{Ridge} = arg\min_{\beta} \frac{1}{n}\left(\epsilon^{'}\epsilon + \lambda I\|\beta\|_2\right) \tag{4.2}$$

where I is $(p \times p)$ identity matrix and by solving the optimization in (4.2) 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's good choice for high dimensions and Big Data analysis.

In terms of matrices, the optimization in (4.2) will be as follows.

$$\begin{aligned} &= (Y - X\beta)^{'}(Y - X\beta) + \lambda I\beta'\beta \\ &= Y^{'}Y - 2\beta' X^{'}Y + \beta^{'}X^{'}X\beta + \lambda I\beta'\beta \end{aligned}$$

By differentiating with respect to β and equalize to zero [6].

$$0 = -2X^{'}Y + 2X^{'}X\beta + 2\lambda I\beta \tag{4.3}$$

$$X^{'}Y = (X^{'}X - \lambda I)\beta \tag{4.4}$$

$$\beta^{Ridge} = (X^{'}X - \lambda I)^{-1}X'Y \tag{4.5}$$

The estimation in (4.5) has been labeled " Ridge Regression". Choosing the complexity parameter λ which is called ridge parameter, has been developed in many approaches some of them suggest to use numerical value $0 < \lambda < 1$ while others suggest to use λ which minimizes the mean of the square errors of the parameters in the model (2.1) so if we express it in the canonical form and let D be an orthogonal matrix, which implies $D^{'}X^{'}XD = \Lambda$ where $\Lambda = diag(k_1, k_2 \ldots k_p)$consist of the eagen vales of $X'X$, then the equivalent regression model will be $Y = X^*\alpha + \epsilon$, where $X^* = XD, \alpha = D'\beta$, here $\alpha = \Lambda^{-1}X^*Y$ represent the OLS estimators for the equivalent model [3, 5].

Hoerl and Kennard 1970 [7] suggest using $\lambda_{HK} = \frac{\sigma^2}{\max \alpha}$ as a ridge parameter where $\sigma^2 = \frac{Y'[I - X(X^{'}X)^{-1}X']Y}{(n-p)}$, this ridge parameter is a suitable to estimate parameters the model (2.1) when there are large sets of data under study.

In 2014 Dorugade [4] suggest a new ridge parameter to use in case of a very large sets of data for the ridge regression estimators in (4.5) as follows.

$$\lambda = \frac{2\sigma^2}{K_{\max}} \sum_{j=1}^{p} \frac{1}{\alpha_j^2}, \tag{4.6}$$

where $K_{\max}$ the maximum value from Eigenvalues for the matrix $X'X$. The formula in (4.5) will be used in our study for the ridge regression method to analyze Big Data sets. The relation between

the Ridge estimates and the Ordinary Least Squares estimate is through the following alternative form so if we multiplies the equation (2.4) by $X'X(X'X)^{-1}$ we can easily get [7]

$$\beta^{Ridge} = (I - \lambda(X^{'}X)^{-1})^{-1}\beta^{OLS} \tag{4.7}$$

$$\beta^{Ridge} = Z\beta^{OLS} \tag{4.8}$$

where $Z = \left(I - \lambda(X^{'}X)^{-1}\right)^{-1}$. In addition, that will lead to the following result.

$$\beta^{Ridge} < \beta^{OLS}. \tag{4.9}$$

The result above constructed upon (4.8) because from definition it is clear that both Z and $X^{'}X$ are symmetric positive definite matrices then ridge regression estimator is better than OLS estimator.

5. Mean Square Errors MSE

Comparison among estimators is very important for any scientific research because it helps the researchers to take the better and the efficient statistical method of analyzing or model selection. Moreover, it helps managers to take decisions. Here we chose the classical Mean Square errors MSE to be as a comparison tool between the two methods OCMT and the Ridge Regression as follows [5].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2 \tag{5.1}$$

In addition, the value of MSE in (5.1) will determine the better method of analyzing as it is small as its better.

6. Data and Results

We have sets of surveys data from the Central Statistical Organization IRAQ represent 300 group of families, and from it, we get the School Dropout rate SDR as follows.

$$SDR = \frac{D}{T} \tag{6.1}$$

Where D is the number of the kids who left school under 12 years age and T is the total kids under 12 years age for each group and this will be vector (300×1) represents our independent variable vector y.

We also have data from for the same groups of families represents many biological and social scales, we get 100 variables from different types quantitative, ordinal, nominal ...etc., as it shows in Appendix. Then these variables construct the matrix X with (200×100) where $n = 300, p = 100$. Estimators is evaluated for both OCMT and Ridge methods according to equations (5),(8) and MSE was calculated according to (13).

From Figure 1, the values of MSE for the both methods, we can notice that the Ridge Estimator is better than the OCMT estimator when the number of variables under study is small but when $p > 75$ we can see that OCMT estimator provides a better act than Ridge estimator. In addition, the difference between the two methods get quite bigger as p goes larger. The figure above make it clear that the OCMT estimator begin with weak values of MSE due the procedure of obtaining it. But it get better as the number of variables grow bigger.

For the Ridge estimator the Figure (1) shows that its good performance with small number of variables in the model .

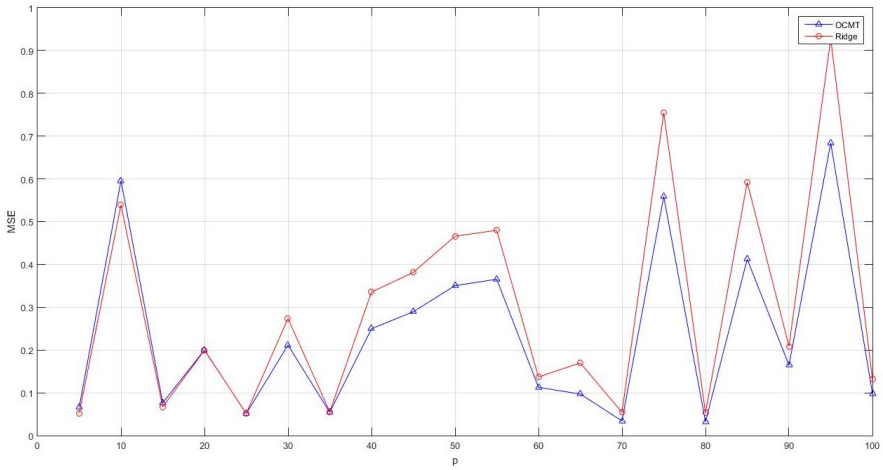

Figure 1: MSE for OCMT and Ridge methods

Table 1: MSE values for both methods

p	Ridge	OCMT	p	Ridge	OCMT
5	0.052	0.0666	50	0.4798	0.3654
10	0.5388	0.5959	55	0.1369	0.928
15	0.0671	0.0765	60	0.1698	0.0971
20	0.1994	0.2001	65	0.0543	0.0342
25	0.0516	0.0512	70	0.7543	0.5588
30	0.2737	0.296	75	0.0531	0.0322
35	0.0555	0.0543	80	0.5921	0.4123
40	0.3354	0.2498	85	0.2084	0.1645
45	0.3813	0.2893	90	0.9273	0.6838
50	0.4659	0.3505	100	0.1315	0.0982

7. Conclusions

From section 6 data and results, we can conclude that the Ridge estimator is better method when the number of variables p are quite small relative to the sample size, and when p grows larger relative to the sample size we can determine that OCMT estimator is remarkably better than Ridge estimator with Big Data conditions.

References

[1] D. Acharjya, A. Kauser, *A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools*, International Journal of Advanced Computer Science and Applications. 2 (2016) 59-518.

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

[3] A. Chudik, G. Kapetanios and M. Pesaran, *One-Covariate at Time, Multiple Testing Approach to variable selection in High-Dimensional Regression Models*, Econometrica, 4 (2018) 1479-1512.

[4] A. Dorugade, *New Ridge Parameters for Ridge Regression*, Journal of the Arab Universities for Basic and Applied Statistics, 3 (2014) 94-99.

[5] J. Fan, H. Fang, *Challenges of Big Data Analysis*, National Science Review, 1 (2014) 293-314.

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

[7] A. Hoerl and A. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems* Technometrics, 1 (1970) 55-67.
[8] 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.
[9] J. Lv, Y. Fan, *A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares*, The Annals of Statistics, 4 (2009) 3498-3528.
[10] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, Jornal of Royal Statistics Society. 1 (1996) 1456-1490.

Appendix A. Demographical Data details.

Table A.2: 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
9	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

Table A.3: N: Nominal - Q: Quantitative - O: Ordered - B: Binary (0,1)

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 edu*	*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*