Journal of Data Science **11**(2013), 323-342

Regression Analysis of Collinear Data using r-k Class Estimator: Socio-Economic and Demographic Factors Affecting the Total Fertility Rate (TFR) in India

Piyush Kant Rai*, Sarla Pareek and Hemlata Joshi
Banasthali University

Abstract:

A basic assumption concerned with general linear regression model is that there is no correlation (or no multicollinearity) between the explanatory variables. When this assumption is not satisfied, the least squares estimators have large variances and become unstable and may have a wrong sign. Therefore, we resort to biased regression methods, which stabilize the parameter estimates. Ridge regression (RR) and principal component regression (PCR) are two of the most popular biased regression methods which can be used in case of multicollinearity. But the r-k class estimator, which is composed by combining the RR estimator and the PCR estimator into a single estimator gives the better estimates of the regression coefficients than the RR estimator and PCR estimator.

This paper explores the multiple regression technique using r-k class estimator between TFR and other socio-economic and demographic variables and the data has been taken from the National Family Health Survey-III (NFHS-III): 29 states of India. The analysis shows that use of contraceptive devices shares the greatest impact on fertility rate followed by maternal care, use of improved water, female age at marriage and spacing between births.

Key words: Multicollinearity, principal component regression (PCR) estimator, r-k class estimator, ridge regression (RR) estimator, total fertility rate (TFR).

1. Introduction

In developing countries, overpopulation is considered to be one of the most basic cause of underdevelopment. The developing countries already facing a lack in their resources, and with the rapidly increasing population, the resources available per person are reduced further, leading to increase poverty, malnutrition, and

*Corresponding author.

other large population related problems. Given this situation, the governments of developing countries, along with non-government organizations, are trying to address this problem by conducting research on the determinants of fertility. India is also dealing with this acute problem, which tends to nullify most of the efforts to encourage development. The government of India has been organizing several programs for controlling the population increase and has been investing the lot of money for controlling the birth rate. Some of the programs have been successful and the rate of increase has also reduced, but has still to reach the sustainable rate. A question of concern to demographers and other social scientists is whether this decline in fertility has been fostered mainly by the family planning programs. Indeed, this reduction in fertility has in some cases led to the belief that the gap between the fertility levels of the different states of India can be substantially reduced by the socialization of family planning services. Available evidence, however, showing that the India has considerable fertility as well as contraceptive use differentials among the various states. These differentials can well be attributed to the fact that socio-economic factors are often differentially distributed across social groups that exists in a society or between societies. Moreover, given that various states of India differ considerably in terms of socio-economic development, it may be that greatest reduction in fertility in those states that experienced significant socio-economic development.

The effect of socio-economic factors on fertility have been examined in a number of studies. Education depresses fertility by increasing the age at marriage, and by increasing the likelihood of contraceptive use (Casteline *et al.*, 1984; Diamond *et al.*, 1997). Other researchers also reported similar depressant effect of education on fertility (Entwisle and Mason, 1985; Rubin-Kurtzman, 1987; Jiang, 1986; Krishnan, 1988; Prada and Ojeda, 1986; Shapiro and Tambashe, 1994; Kravdal, 2002). Place of residence has also been found to be significantly related to fertility: total fertility rates are higher among rural women than among urban women (Alam and Casterline, 1984; Rubin-Kurtzman, 1987; Prada and Ojeda, 1986). Income is negatively related to fertility (Rubin-Kurtzman, 1987; Jiang, 1986).

Since Total Fertility rate is the most important measure of fertility in demography and TFR is affected by many socio-economic and other development factors. Hence the main objective of this paper is to know that up to what extent and how the socio-economic and other development factors impact the fertility level of India. It is believed that socio-economic and other development factors do exert significantly independent as well as the joint impact on fertility after eliminating the effect of the family planning programs and policies. An attempt has been made in this paper to identify these factors and their relative contributions towards the variation in the fertility level of India. The importance of the

study derives from the fact that it is necessary to identify those population groups whose fertility is high but reducible through changes in government policies and the redistribution of available resources.

2. Data and Method

The dependent variable is the total fertility rate (defined as average number of children a woman has in her lifetime). Total fertility rate is affected by many demographic, social, cultural, and economic variables. The explanatory variables considered in the present study are those, that appeared influential in fertility variation. These variables are Human development index (HDI = X_1), infant mortality rate (IMR = X_2), defined as infant deaths per thousand live births, percent of population using contraceptive devices (any method = X_3), median age at marriage of male (= X_4), median age at marriage for female (= X_5), median number of months since preceding birth (= X_6), percent of population using improved water for drink (= X_7), male literacy rate (= X_8), female literacy rate (= X_9) and percent of mothers who are taking maternal care (= X_{10}). Here, the independent variables that we have considered are discrete as well as of continuous in nature for e.g., X_4, X_5 and X_6 are continuous variables while the others are discrete in nature. But we have confined ourselves for the integral values of the age at marriage and birth intervals. But it can be considered as the continuous case (Sufian, 2005). In the analysis, the data on the several variables was taken from National Family Health Survey-III (NFHS-III) about 29 Indian states. National family health survey is the nationwide sample survey which consider the following sampling design and techniques of data collection.

Sample Design: The urban and rural samples within each state were drawn separately and, to the extent possible, the sample within each state was allocated proportionally to the size of the state's urban and rural populations. A uniform sample design was adopted in all the states. In each state, the rural sample was selected in two states: the selection of primary sampling units (PSUs), which are villages, with probability proportional to population size (PPS) at the first stage, followed by the random selection of households within each PSU in the second stage. In urban areas, a three-stage procedure was followed. In the first stage, wards were selected with PPS sampling. In the next stage, one census enumeration block (CEB) was randomly selected from each sample ward. In the final stage, households were randomly selected within each sample CEB. Each ward comprises several enumeration blocks (CEB) created for the census. A list of all the CEBs in a selected ward formed the sampling frame at the second stage. Such lists of CEBs in the selected wards were made available for use for NFHS-III by the census office on request. Each CEB is comprised of about 150-200 households.

Sample Selection: In rural areas, the 2001 Census list of villages served as the sampling frame. the list was stratified by a number of variables. The first level of stratification was geographic, with districts being subdivided into contiguous regions. Within each of these region, villages were further stratified using selected variables from the following list: village size, percentage of males working in the non-agricultural sector, percentage of the population belonging to scheduled castes or scheduled tribes, and female literacy. In addition to these variable, HIV prevalence status, i.e., "High", "Medium" and "Low" as estimated for all the districts in high HIV prevalence states, was used for stratification in the high HIV prevalence states. Female literacy was used for implicit stratification (i.e., the villages were ordered prior to selection according to the proportion of females who were literate) in most states although it may be an explicit stratification variables in a few states.

The mean and standard deviation are given in the table below.

Table 1: Means and standard deviations of total fertility rate and ten predictor variables: 29 states of India

Variables	Descriptive Statistics	
	Mean	Standard deviation
TFR	2.628	0.695
HDI	0.629	0.112
IMR	47.272	15.881
Contraceptive Use	55.121	12.815
Male age marriage	22.231	12.142
Female age at marriage	35.434	15.792
Birth Interval	32.441	3.139
Improved Water	81.934	12.358
Male Literacy rate	80.817	7.798
Female Literacy rate	62.655	15.674
Maternal Care	60.421	20.284

Table 1 presents the means and standard deviations of the dependent as well as explanatory variables. State-wise TFR is taken in data and then the mean and standard deviation has been computed. The TFR has an average value of 2.62 children per woman varying from lows of 1.79 children in Goa and Andhra Pradesh, 1.8 children in Tamil Nadu, 1.94 children in Himachal Pradesh and many other states also highs of 4.2 children in Bihar, 3.8 children in Meghalaya and 3.6 children in Jharkhand. We hypothesize that contraceptive use, female and male age at marriage, birth interval, use of improved water, HDI, female and male literacy rate and maternal care will be negatively related to the total fertility rate while positive relationships are expected between TFR and each of the IMR.

The most commonly used estimator for the estimation of parameters is the ordinary least square (OLS) estimator. Under certain assumptions, least square method produce estimators with desirable properties. In some instances (e.g., when one or more assumptions do not hold) other estimators may be superior to ordinary least square (OLS). The other estimators are maximum likelihood, ridge, principal components and r-k class estimator.

The "n" observations for the dependent variable Y are determined by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots n, \tag{2.1}$$

(2.1) can also be written as

$$Y = \beta_0 + \sum_{j=1}^{10} \beta_j X_j + \varepsilon, \tag{2.2}$$

where Y is the response variable, i.e., TFR and $X_1, X_2, \cdots, X_{10}$ are the predictor variables, β_0 is the intercept term. It gives the mean or average effect on Y of all the variables excluded from the model and βi's are partial regression coefficients or the slope parameters describing the relation between the response and predictor variables, on the other hand partial regression coefficients measures the change in the mean value of Y corresponding to per unit change in X_j, when all other predictor variables are held constant.

Consider the standard matrix form of the above multiple linear regression model

$$Y = X\beta + \varepsilon, \tag{2.3}$$

where $X = (x_{ij})$ is a fixed $n \times p + 1$ matrix. [(x_{ij}) is the i^{th} observation on the j^{th} independent variable] and is of full rank p ($p \leq n$), $Y = (y_i)$ is an $n \times 1$ vector of observations on the dependent variables, β is a $p + 1 \times 1$ unknown column vector of regression coefficients, and $\varepsilon = (\varepsilon_i)$ is an $n \times 1$ vector of random errors; $E(\varepsilon) = 0, E(\varepsilon\varepsilon') = \sigma^2 I_n$, where I_n denotes the $n \times n$ identity matrix, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Let us assume that the variables have been standardized by subtracting their sample means and dividing by their sample standard deviations. Then the model given in (2.3) will be

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}. \tag{2.4}$$

Now, we wish to estimate the $p \times 1$ vector β of regression coefficients. The variables are assumed to be standardized so that $X'X$ is in the form of correlation matrix, and the vector $X'Y$ is the vector of correlation coefficients of the dependent variable with each explanatory variable.

The least squares (LS) estimator, $\hat{\beta}$ of the parameters are given by

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.5}$$

Here the assumption for ordinary least square (OLS),

1. X is set of fixed numbers.
2. X is full column rank matrix, i.e., rank of X should be p.
3. Predictor variables $(X_1, X_2, \cdots, X_p)$ are linearly independent, i.e., $X'X$ is non-singular matrix means $|X'X| \neq 0$.

OLS has been treated as the best estimator for a long time. However, many results have proved that the OLS estimator is no longer a good estimator when the multicollinearity is present (Al-Hassan, 2008). In multiple linear regression models, we usually assume that the explanatory variables are independent. However, in practice, there may be strong or near to strong linear relationships among the explanatory variables. In that case the independent assumptions are no longer valid, which causes the problem of multicollinearity. Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly linearly related or correlated. If our goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will still be accurate, and the overall R^2 quantifies how well the model predicts the Y values. If our goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. In the presence of multicollinearity, it is impossible to estimate the unique effects of individual variables in the regression equation. Multicollinearity increases the standard errors of the coefficients. Increased standard error means that coefficients for some independent variables may be found insignificant, whereas without multicollinearity and with lower standard errors, these same coefficients might have been found to be significant. Moreover, the LS estimates are likely to be too large in absolute value and possibly, of the wrong sign (Al-Hassan, 2008). Therefore, multicollinearity becomes one of the serious problems in the linear regression analysis. Multicollinearity only affects calculations regarding individual predictors, i.e., a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

Diagnosing Multicollinearity

In order to lay the foundation for detection of multicollinearity problem, some classic symptoms are present in our data:

- The F is highly significant (p-value-0.000), implying that the variables are chosen are valid explanatory variables (Chatterjee and Price, 1977, p. 146) and most of the regression coefficients are insignificant at 5% level of significance, which can be seen from the Table 2.
- The value of R^2 is quite large, i.e., 0.899.
- Variance inflation factor (VIF) of HDI and Female literacy rate is greater than 10, i.e., 10.337 and 11.019 respectively.
- Sometimes eigenvalues, condition indices and the condition number can be referred in examining multicollinearity. The condition number (k) is given as the square root of the largest eigenvalue ($\max(\lambda)$) divided by the smallest eigenvalue ($\min(\lambda)$), i.e.,

$$k = \sqrt{\frac{\max(\lambda)}{\min(\lambda)}}.$$

  In our case, $k = 11.406$, when there is no collinearity at all, the eigenvalues, condition indices and condition number will all equal to one. As collinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem), and the condition indices or the condition number will increase.

Table 2: β-Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	p-value	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
Constant	4.679	1.975		2.370	0.029		
HDI	0.630	1.490	0.102	0.423	0.678	0.097	10.337
IMR	-0.001	0.010	-0.029	-0.131	0.898	0.113	8.857
Contraceptive Use	-0.034	0.007	-0.620	-4.705	0.000	0.322	3.101
Male age marriage	-0.008	0.011	-0.136	-0.698	0.494	0.147	6.787
Female age at marriage	0.020	0.008	0.445	2.314	0.033	0.152	6.587
Birth Interval	-0.025	0.024	-0.111	-1.019	0.322	0.472	2.117
Improved Water	-0.007	0.006	-0.121	-1.081	0.294	0.444	2.253
Male Literacy rate	0.013	0.018	0.144	0.706	0.489	0.135	7.402
Female Literacy rate	0.005	0.011	0.117	0.472	0.643	0.091	11.019
Maternal Care	-0.018	0.006	-0.511	-2.785	0.012	0.167	6.002

$N = 29, R^2 = 0.899$, adjusted $R^2 = .843$ and $F = 16.04$

From the Table 3, we can see that how the explanatory variables are correlated. Among the explanatory variables, HDI and IMR are highly negatively correlated (-0.89), correlation between HDI and Female age at marriage is -0.74 and there is high positive correlation between the female literacy rate and HDI (0.89) and similarly, we can see the correlation between the all other predictor variables. A correlation is the measurement of the relationship between two variables. A positive correlation is a direct relationship, as the amount of one variable increases, the amount of a second variable also increases. And in a negative correlation, as the amount of one variable goes up, the levels of another variable goes down.

Table 3: Correlation matrix of predictor variables

Variables	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$	$X10$
$X1$	1.00									
$X2$	-0.89	1.00								
$X3$	0.34	-0.31	1.00							
$X4$	-0.66	0.77	-0.34	1.00						
$X5$	-0.74	0.80	-0.15	0.77	1.00					
$X6$	0.35	-0.36	0.25	-0.53	-0.26	1.00				
$X7$	-0.07	0.03	0.50	0.11	0.20	-0.34	1.00			
$X8$	0.86	-0.81	0.42	-0.64	-0.80	0.26	0.03	1.00		
$X9$	0.89	-0.86	0.35	-0.79	-0.84	0.40	-0.09	0.86	1.00	
$X10$	0.65	-0.76	0.46	-0.82	-0.60	0.54	-0.07	0.53	0.64	1.00

One of the other way to check the multicollinearity is that if sum of the reciprocals of the eigenvalues is greater than five times of the number of predictor variables used then there is multicollinearity in the data. And in this data, sum of reciprocals of the eigenvalues is 64.46 which is greater than five times the number of predictor variables (10) used [47]. All these indicates the presence of multicollinearity. And in case of presence of multicollinearity, the estimates obtained by OLS estimator are not reliable and desirable if we want to know that how predictor variables (X) impacts on response variable (Y).

Several methods have been suggested to solve this problem. Ridge regression (RR) and principal component regression (PCR) are two of the most popular biased regression methods that help to discuss the problem of collinearity in the data and provide the better solution of the problem.

1. Ridge Regression (RR): Hoerl and Kennard (1970a) suggested the use of $X'X + kI_p, (k \geq 0)$ rather than $X'X$, in the estimation of β (2.5). The resulting estimators of β are known in literature as the RR estimator, given by

$$\hat{\beta(k)} = (X'X + kI_p)^{-1}X'Y. \tag{2.6}$$

The constant k is known as biasing or ridge parameter. As k increases from zero and continues up to infinity, the regression estimates tend toward zero. Though these estimators result in bias, for certain value of k, they yield minimum MSE compared to the LS estimator (Hoerl and Kennard, 1970a). However, the MSE($\hat{\beta}(k)$) will depend on unknown parameters k, β and σ^2, which cannot be calculated in practice. But k has to be estimated from the real data instead. Several methods for estimating k have been proposed and evaluated by several researchers. Some of these researchers are Hoerl and Kennard (1970a), Hoerl *et al.* (1975), McDonald and Galarneau (1975), Lawless and Wang (1976), Hocking *et al.* (1976), Wichern and Churchill (1978), Nordberg (1982), Saleh and Kibria (1993), Singh and Tracy (1999), Wencheko (2000), Kibria (2003), Khalaf and Shukur (2005) and Al-Hassan (2008). But Geometric Mean Method ($\hat{k}_{\text{GM}}$ or GM) performs better than the other estimators when the correlations between the explanatory variables are moderate, and for high correlations, HKB becomes better than GM, and for extremely high correlation all estimators (except AM) perform better than or as good as GM (Al-Hassan, 2008).Thus out of these methods for estimating k, Hoerl, Kennard and Baldwin Method ($\hat{k}_{\text{HKB}}$ or HKB) and Geometric Mean Method ($\hat{k}_{\text{GM}}$ or GM) gives the better estimate of k than the others. And in our paper, there are almost moderate correlation between the explanatory variables. Thus we have used Geometric Mean Method ($\hat{k}_{\text{GM}}$ or GM) for estimation of k.

   (i) Hoerl, Kennard and Baldwin Method ($\hat{k}_{\text{HKB}}$ or HKB): Hoerl, Kennard and Baldwin proposed a different estimator of k than the others by taking the harmonic mean of $\hat{k}$. That is

$$\hat{k}_{\text{HKB}} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\hat{\beta}^2}. \tag{2.7}$$

   (ii) Geometric Mean Method ($\hat{k}_{\text{GM}}$ or GM): Kibria proposed estimating k by using the geometric mean of $\hat{k}$, which produces the following estimator

$$\hat{k}_{\text{GM}} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p}\hat{\beta}^2)^{\frac{1}{p}}}. \tag{2.8}$$

2. Principal Component Regression (PCR): In principal component analysis, the p original variables are transformed into linear combinations called principal components. Principal components were first proposed by Person (1901) and further developed by Hotelling (1933). Comprehensive surveys of the field have been given by Jolliffe (1986), Jackson (1991) and Basilevsky (1994). Other reviews are by Rao (1964), Jackson (1980; 1981), Wold *et al.* (1987), Duntman (1989) (Rencher, 1998) and Jolliffe (2005). As we have indicated, an approach to the problem of multicollinearity is PCR, in which Y is regressed on the principal components of X's . If we use only the larger principal components, the large variances in $\hat{\beta}_j$'s due to multicollinearity are reduced, but of course we introduce some bias in the new $\hat{\beta}_j$'s. Often, the principal components with the highest variance are selected. However, the low variance principal components may also be important, and in some cases, they may even more important than those with the highest variances (Jolliffe, 1982).

   If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be the eigenvalues of the correlation matrix of explanatory variables and $e_1, e_2, \cdots, e_p$ are orthogonal eigenvectors corresponding to the eigenvalues. Orthogonal means

$$e_i' e_j = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Let T be the $(p \times p)$ orthogonal matrix, i.e., $T = (t_1, t_2, \cdots, t_p)$ such that it diagonalizes $X'X$, i.e., $T'X'XT = \Lambda = \text{diag}(e_1, e_2, \cdots, e_p)$ where $T'T = Ip = TT'$, being diagonal matrix consisting of eigenvalues of $X'X$ as its diagonal elements. (2.4) can be written as

$$\begin{aligned} Y &= XTT'\beta + \varepsilon \\ \Rightarrow Y &= (XT)(T'\beta) + \varepsilon \\ \Rightarrow Y &= X^*\alpha + \varepsilon, \end{aligned}$$

where $X^* = (XT)$ and $\alpha = (T'\beta)$

$$\begin{aligned} \hat{\alpha} &= (X^{*'}X^*)^{-1}X^{*'}Y, \\ T'\beta &= (T'T'XT)^{-1}T'X'Y, \end{aligned}$$

on pre-multiplying by T both sides, we have

$$\begin{aligned} TT'\beta &= T(T'X'XT)^{-1}T'X'Y, \\ \hat{\beta} &= T(T'X'XT)^{-1}T'X'Y. \end{aligned} \tag{2.9}$$

After deleting $(p-r)$ columns of T, T_r be the remaining eigenvectors of $X'X$ so that $T_r^{'}X'XT_r = \Lambda_r$ then from (2.9), the reduced model will be

$$\hat{\beta}_r = T_r(T_r^{'}X'XT_r)^{-1}T_r^{'}X'Y, \tag{2.10}$$

here T_r will be of $p \times r$ matrix of eigenvectors.

The purpose of principal components is to generate a reduced set of variables that account for most of the variance of the original variables. We must therefore decide just how many components to retain; other components will be discarded. In reality, the number of components extracted in a principal component analysis is equal to the number of observed variables being analysed. However, Mansfileld *et al.* (1977) suggested that only the first few components account for meaningful amounts of variance, so only these first few components are retained and used in multiple regression analyses. Jolliffe (1982) represents the point of view of many statisticians whose decisions depend only on the magnitude λ of the variance of the principal component.

The eigenvalues of the correlation matrix of the predictor variables have also been calculated. Which are given by

$$\lambda_1 = 5.9718, \lambda_2 = 1.5576, \lambda_3 = 1.0881, \lambda_4 = 0.4742, \lambda_5 = 0.2966,$$

$$\lambda_6 = 0.2705, \lambda_7 = 0.1258, \lambda_8 = 0.0929, \lambda_9 = 0.0766, \lambda_{10} = 0.0459.$$

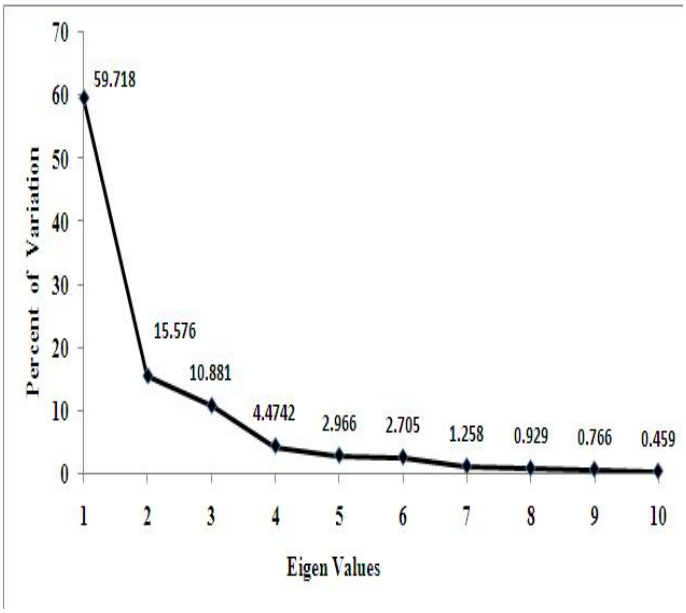

Figure 1: Scree-plot

The above figure shows the percent of variation explained by the principle components. From this figure we can conclude that first principle component is showing 59.7% variation, second principle component is showing 15.56% and so on. First four principle components are showing 90.91% variation therefore we will consider only first four principle components in our model. The first four eigenvalues are $\lambda_1 = 5.9718, \lambda_2 = 1.5576, \lambda_3 = 1.0881$ and $\lambda_4 = 0.4742$ and the corresponding eigenvectors are given in the matrix T_r below:

$$T_r = \begin{array}{c} \begin{matrix} e_1 & e_2 & e_3 & e_4 \end{matrix} \\ \begin{pmatrix} 0.37 & 0.01 & -0.19 & 0.3 \\ -0.38 & -0.01 & 0.14 & 0.13 \\ 0.18 & 0.58 & 0.37 & 0.26 \\ -0.36 & 0.07 & -0.17 & 0.45 \\ -0.35 & 0.14 & 0.29 & 0.19 \\ 0.21 & -0.25 & 0.68 & 0.36 \\ -0.03 & 0.75 & -0.06 & -0.21 \\ 0.36 & 0.11 & -0.28 & 0.37 \\ 0.38 & -0.02 & -0.16 & 0.18 \\ 0.34 & 0.01 & 0.34 & -0.49 \end{pmatrix} \end{array}.$$

And by using the Geometric mean method given in (2.8), the value of k is calculated, i.e., $k = 2.723501009$.

PCR was first proposed by Hotelling (1957) and Kendal (1957). Hsuan (1981) explored the relationship between PCR and RR. He proved that when the data are severely multi-collinear, the ridge estimator can be made very close to the principal components estimators. Baye and Parker (1984) and Nomura and Ohkubo (1985) proposed the $r - k$ class estimator by combining the RR estimator and the PCR estimator into a single estimator, which performs better than the other estimators while dealing with multicollinearity (Sarkar, 1989). The $r - k$ class estimator can be written in the form:

$$\hat{\beta}_r(k) = T_r(T_r'X'XT_r + kI)^{-1}T_r'X'Y, \tag{2.11}$$

where T_r is the matrix of eigenvectors, X and Y is the standardized matrices of explanatory and response variables respectively and I is the $(r \times r)$ identity matrix.

Now from the above $r-k$ class estimator given in (2.11), we can easily estimate the regression coefficients. Which are given in the Table 4 below:

Table 4: Regression coefficients calculated by using r-k class estimator

Variables	regression coefficients	t	p-value
HDI	0.0176	0.2351	0.5916
IMR	0.0714	0.9536	0.8235
Contraceptive use	-0.3386	-4.5222	0.0001*
Male age at marriage	0.0325	0.4341	0.6653
Female age at marriage	-0.2071	-2.7659	0.0064*
Birth Interval	-0.1612	-2.1529	0.0226*
Use of improved water	-0.2582	-3.4483	0.0014*
Male literacy rate	0.0330	0.4407	0.6677
Female literacy rate	-0.0073	-0.0975	0.4617
Maternal care	-0.2968	-3.9639	0.0005*

*p-value significant (< 0.05)

Finally, our regression model can be written as

$$Y = 0.0176X_1 + 0.0714X_2 - 0.3386X_3 + 0.0325X_4 - 0.2071X_5 - 0.1612X_6 - 0.2582X_7 + 0.0334X_8 - 0.0073X_9 - 0.2968X_{10}.$$

Or we can write

$$\begin{aligned}\text{Total Fertility Rate} = {} & 0.0176(\text{HDI}) + 0.0714(\text{IMR}) - 0.3386(\text{Contraceptive use}) \\ & +0.0325(\text{Male age at marriage}) - 0.2071 \\ & -0.1612(\text{Birth Interval}) - 0.2582(\text{Use of Improved water}) \\ & +0.0334(\text{Male literacy rate}) - 0.0073(\text{Female literacy rate}) \\ & -0.2968(\text{Maternal Care}). \end{aligned} \tag{2.12}$$

3. Conclusion

India, being a developing country, has to face several socio-demographic challenges. One of the most important problem is the population explosion or the high birth rate. There are lot of problems associated with high birth rate. High birth rates can cause stress on the government welfare and family programs to support a youthful population. Additional problems faced by a country with a high birth rate include educating a growing number of children, creating jobs for these children when they enter to the workforce, and dealing with the environmental effects that a large population can produce. Several solutions to decrease the rate of population increase has been tried by the government of India, some successfully, some unsuccessfully. Although the rate of increase has decreased up to some extent but the rate has not reached to the satisfactory level yet. The

population of India continues to increase at an alarming rate. The effects of this population increase are evident in the increasing poverty, unemployment, air and water pollution, shortage of food, health sources and educational resources. Thus it is important to analyse the determinants of fertility in India to identify their relative weights necessary for ascertaining priorities while formulating population policies.

In order to evaluate the relative importance of the explanatory variables in determining total fertility rate, the standardized variables have been used. The resulted regression model given in (2.12) support the conclusion that use of contraceptive devices (any method) is very useful factor that has the highest impact to decrease the total fertility rate. The family planning policies and programs are the most important contributor in reduction of fertility rate. This lends support to the contention that the determinative factor that has fostered the recent decline in fertility in India has been mainly by the government's family planning programs.

The other variable significantly related to the total fertility rate is the maternal care. A healthy, relaxed mother would be more likely to have a positive effect on the well-being of the new born. If there is no care of mother then obviously the child will be very weak and the chances of the infant mortality will be higher. And the need of children make the higher fertility rate. Many studies have obtained results supportive of the positive effect of infant and child mortality on fertility (Adlakha, 1973; Taylor, Newman and Kelly, 1976). The idea is conceptually related to the child survival hypothesis. Experience with, or fear of infant and child mortality might make married couples have extra births to replace young children who already died. As such, societies with higher infant mortality tend to have higher fertility. Thus, the overall purpose to reduce the fertility rate is to make an improvement in mother and child health.

In developing country like India, the use of improved water play a very important role in the human fertility. Here in our case, its role looks vital for deciding the TFR for the women in the period of reproductive age. We see that TFR decreases with the use of improved water slightly more than the other negative factors like female age at marriage, birth interval and female literacy rate. Thus it is quite interesting to analyse the role of use of improved water on human fertility, which is very important factor for civilize society.

The fourth important variable known to influence the fertility performance of women is the female age at marriage, in the sense that if the female age at marriage is low, women start having their children at an early age, and these children, in their turn, begin to procreate early. By rising the age at marriage, specially for women, we cut down on their reproductive span and thus reduce fertility.

The role of education is widely believed to be central to major changes in fertility rate in India and elsewhere. Generally, having a higher level of education is associated with later and less childbearing and higher-educated women are more likely to have higher earning husbands or partners, so providing a further positive "income effect" on childbearing. Education also provides an opportunity to participate in gainful employment outside the home, and this competes with the demands of childbearing. Better educated women enjoy better access to opportunities of life, and hence lower fertility is felt more advantageous to them than higher fertility, since with lower fertility it is easier to reap the benefits of those opportunities. Thus an educated woman is very likely to prefer a smaller family. Among women with no education even significant difference in the number of children fails to make any observable difference in the level of living. As such, societies with lower level of literacy have greater likelihoods of having larger fertility rates. Education exposes a woman to a wide range of information regarding birth control and family planning and decreases the total fertility rate.

Birth interval also impact the fertility. The model says that birth interval is also the factor that decreases the total fertility rate but at very low level. This indicates that in India, the spacing between the two birth is still low.

All these factors have implications for their fertility performance. Thus, although the family planning programs have played the most important roles in declining fertility, this decline should not be viewed as due, solely, to successful family planning programs. The results of this analysis indicate that an egalitarian distribution of the benefits of socio-economic development over rural and urban areas, maternal care, an increase in the level of female literacy, decrease in the level of infant mortality, use of improved water for drink, age at marriage and spacing between birth may be important strategies for reducing the fertility rates in India. But the raising of age at marriage will have an impact on fertility only when the law relating to it is uniformly enforced throughout the country.

Acknowledgements

The authors are thankful to the Editor-in-Chief Professor Dr. Wen-Jang Huang and learned referees for bringing the original manuscript into its present form. The author would also like to thank Dr. Shalini Chandra Dept. of Mathematics and Statistics, Banasthali University, Rajasthan, India for her valuable suggestions and help in the manuscript.

References

[1] Adlakha, A. L. (1973). Fertility and infant mortality: an analysis of Turkish data. *Demography India* **2**, 56-76.

[2] Alam, I. and Casterline, J. B. (1984). Socioeconomic differentials in recent fertility. International Statistical Institute. *World Fertility Survey Comparative Studies*: Cross-National Summaries, No.33. Voorburg, Netherlands.

[3] Al-Hassan, Y. M. (2008). A Monte carlo evaluation of some ridge estimators. *Jordan Journal of Applied Science* **10**, 101-110.

[4] Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician* **27**, 17-21.

[5] Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*: *Theory and Applications.* Wiley, New York.

[6] Caldwell, J. C., Phillips, J. F. and Barkat-e-Khuda. (2002). Family planning programs in the twenty-first century. *Studies in Family Planning* **33**, 1-10.

[7] Casterline, J. B., Singh, S., Cleland, J. and Ashurst, H. (1984). The proximate determinants of fertility. International Statistical Institute. *World Fertility Survey Comparative Studies*: Cross-National Summaries, No.39. Voorburg, Netherlands.

[8] Chatterjee, S. and Price, B. (1977). *Regression Analysis by Examples.* Wiley, New York.

[9] Chang, X. and Yang, H. (2012). Combining two-parameter and principal component regression estimators. *Statistical Papers* **53**, 549-562.

[10] Cutright, P. and Kelly, W. R. (1981). The role of family planning programs in fertility declines in less developed countries, 1958-1977. *International Family Planning Perspectives* **7**, 145-151.

[11] Diamond, I., Tonkin, P., Rahman, A. P. M. and Noor, S. A. (1997). Spatial variation in contraceptive method use in Bangladesh. In *Bangladesh Demographic and Health Survey 1993-94, Extended Analysis* (Edited by A. Kantner, A. Al-Sabir and N. Chakraborty), 136-157. National Institute of Population Research and Training, Dhaka.

[12] Duntman, G. H. (1989). *Principal Components Analysis*: *Quantitative Applications in the Social Sciences.* Newbury Park, Sage, California.

[13] Entwisle, B. and Mason, W. M. (1985). What has been learned from the World Fertility Survey about the effects of socioeconomic position on reproductive behavior. Population Studies Center Research Report No.85-77. University of Michigan, Population Studies Center, Ann Arbor, Michigan.

[14] Hocking, R. R., Speed, F. M. and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics* **18**, 425-437.

[15] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

[16] Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulation. *Communications in Statistics - Theory and Methods* **4**, 105-123.

[17] Jackson, D. N. and Chan, D. W. (1980). Maximum likelihood estimation in common factor analysis: a cautionary note. *Psychol Bull* **88**, 502-508.

[18] Jackson, J. E. (1991). *A User's Guide to Principal Components.* Wiley, New York.

[19] Jollife, I. T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

[20] Jolliffe, I. T. (2005). *Principal Component Analysis.* In *Encyclopedia of Statistics in Behavioral Science* **3**, 1580-1584. Wiley, New York.

[21] Jiang, Z. (1986). Impact of socioeconomic factors on China's fertility. *Population Research* **3**, 9-17.

[22] Khalaf, G. and Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods* **34**, 1177-1182.

[23] Kibria, B. M. (2003). Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation* **32**, 419-435.

[24] Kravdal, O. (2002). Education and fertility in sub-Saharan Africa: individual and community effects. *Demography* **39**, 233-250.

[25] Krishnan, V. (1988). Homeownership: its impact on fertility. Population Research Laboratory Discussion Paper No.51. University of Alberta, Department of Sociology, Edmonton, Canada.

[26] Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods* **5**, 307-323.

[27] Lewis-Beck, M. S. (1980). *Applied Regression*: *An Introduction.* Sage Publications. Thousand Oaks, California.

[28] Mauldin, W. P. and Berelson, B. (1978). Conditions of fertility decline in developing countries. *Studies in Family Planning* **9**, 89-147.

[29] McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* **70**, 407-412.

[30] Muniz, G. and Kibria, B. M. G. (2009). On some ridge regression estimators: an empirical comparisons. *Communications in Statistics - Simulation and Computation* **38**, 621-630.

[31] Nordberg, L. (1982). A procedure for determination of a good ridge parameter in linear regression. *Communications in Statistics - Simulation and Computation* **11**, 285-309.

[32] Population Reference Bureau, Inc. (1987). *World Population Data Sheet.* Washington, District of Columbia.

[33] Population Information Program. (1985). Population Reports. Series M., Number 8 (Special Topics) **13**, 290. The Johns Hopkins University, Maryland.

[34] Poston, D. L. and Gu, B. (1987). Socioeconomic development, family planning, and fertility in China. *Demography* **24**, 531-551.

[35] Prada, E. and Ojeda, G. (1987). Selected findings from the demographic and health survey in Colombia, 1986. *International Family Planning Perspectives* **13**, 116-120.

[36] Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, Series A* **26**, 329-358.

[37] Rencher, A. C. (1995). *Methods of Multivariate Analysis.* Wiley, New York.

[38] Ross, J. A., Rich, M., Molzan, J. P. and Pensak, M. (1988). *Family Planning and Child Survival*: *100 Developing Countries.* Center for Population and Family Health, Columbia University, New York.

[39] Rubin-Kurtzman, J. R. (1987). *The Socioeconomic Determinants of Fertility in Mexico*: *Changing Perspectives.* Center for U.S.-Mexican Studies Monograph Series No.23. University of California, La Jolla, San Diego.

[40] Saleh, A. K. and Kibria, B. M. (1993). Performances of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics - Theory and Methods* **22**, 2747-2764.

[41] Sarkar, N. (1989). Comparisons among some estimators in misspecified linear models with multicollinearity. *Annals of The Institute of Statistical Mathematics* **41**, 717-724.

[42] Shapiro, D. and Tambashe, B. O. (1994). The impact of women's employment and education on contraceptive use and abortion in Kinshasa, Zaire. *Studies in Family Planning* **25**, 96-110.

[43] Singh, S. and Tracy, D. S. (1999). Ridge-regression using scrambled responses. *Metrika* **41**, 147-157.

[44] Wencheko, E. (2000). Estimation of the signal-to-noise in the linear regression model. *Statistical Papers* **41**, 327-343.

[45] Wichern, D. and Churchill, G. (1978). A comparison of ridge estimators. *Technometrics* **20**, 301-311.

[46] Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 37-52.

[47] Sufian, M. A. J. (2005). Analyzing collinear data by principal component regression approach − an example from developing countries. *Journal of Data Science* **3**, 221-232.

[48] Taylor, C. E., Newman, J. S. and Kelly, N. U. (1976). The child survival hypothesis. *Population Studies* **30**, 263-271.

[49] Tsui, A. O. and Bogue, D. J. (1978). Declining world fertility: trends, causes, implications. *Population Bulletin* **33**, 1-42.

Received June 26, 2012; accepted December 7, 2012.

Piyush Kant Rai
Department of Mathematics and Statistics
Banasthali University
Banasthali Vidyapith, P.O. Banasthali Vidyapith, Rajasthan - 304022, India
raipiyush5@gmail.com

Sarla Pareek
Department of Mathematics and Statistics
Banasthali University
Banasthali Vidyapith, P.O. Banasthali Vidyapith, Rajasthan - 304022, India
psarla13@gmail.com

Hemlata Joshi
Department of Mathematics and Statistics
Banasthali University
Banasthali Vidyapith, P.O. Banasthali Vidyapith, Rajasthan - 304022, India
kirtidbest@gmail.com