IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 9, Issue 6 (Jan. 2014), PP 61-74
www.iosrjournals.org

Comparison of Classical Least Squares (CLS), Ridge and Principal Component Methods of Regression Analyses using Gynecological Data

Olawuwo, Simeon[1]; Ogunleye, Timothy A.[2*]; Ojo, Thompson O[3] & Adejumo, Adebowale O[4]

[2*]*(Department of Budget, Planning, Research & Statistics, Ejigbo Local Government, State of Osun, Nigeria)*
[1,3]*(Department of Statistics, Federal Polytechnic Ede, Private Mail Bag 231, State of Osun, Nigeria)*
[4]*(Department of Statistics, University of Ilorin, Private Mail Bag 1515, Ilorin, Kwara State, Nigeria)*
*2*Corresponding Author's Contacts- Mobile: +23468998580, Electronic mail: thompsondx@gmail.com*

Abstract: *This study compares results obtained from the application of Classical Least Squares with that obtained from the two major biased estimation methods: Ridge and Principal Component Regressions in multicollinear situations using gynecological data from University College Hospital, Ibadan, Oyo State, Nigeria. Numerical values of baby's weights (less than 2.5kg) at birth were considered as response variable while mother's age, weight and height, as well as preterm delivery, multiple pregnancies, parity, graphidity, maternal infections such as malaria, tuberculosis, sexually transmitted diseases, anaemia/shortage of blood, intra-uterine infections, congenital abnormalities, etc and fetal infections serve as explanatory variables. Regression method is used as the statistical tool. A number of assumptions of the regression analysis were inspected. Normality assumption was confirmed by plotting Normal Q-Q Plot and Histogram of the Standardized Residuals. The data sets were also inspected for homoscedasticity of error variances using Residual Plot and Fligner-Killeen Test; it was established that homoscedasticity assumption was not violated. Autocorrelation problem on the data set was checked by Durbin-Watson statistic. The test revealed that autocorrelation problem can be tolerated. Existence of multicollinearity problem was further checked in the data set using Farrar-Glauber Chi-squared test. It was established that some predictors are highly correlated; this was also true when correlation matrix table was obtained. Shrinkage estimator of Ridge Regression was obtained by Iterative Method of Hoerl and Kennard. Ridge Regression coefficients were later computed. Kaiser's and Cattell's Screen Criteria were employed for determining number of principal components to be retained in the analysis. The two criteria suggested that the first three components should be retained but only one component is significant. Thereafter, Ridge Regression was finally recommended as the best method to handle multicollinearity problem under Frequentist Approach.*

Keywords: *Classical Least Squares, Ridge and Principal Component Regressions, Normal Q-Q Plot, Fligner-Killeen and Farrar-Glauber Tests, Durbin-Watson statistic and Shrinkage estimator.*

I. Introduction:

Every parent hopes for a healthy baby and a newborn's weight is an excellent indicator of it. Weight at birth is a good indicator not only of a mother's health and nutritional status but also of the newborn's chances for survival, growth, long-term health and psychological development. The incidence of Low Birth Weight, especially in Nigeria, varies from about (5-40) percent of live births (Martin, 2007). It's been reported that about one-third of the infants weigh less than 2.5kg, which is the mean birth weight of all the newborn babies (Berghella, 2007).

Low Birth Weight (LBW) is a major determinant of infant mortality and morbidity. In fact, it is considered the single most important predictor of infant mortality, especially of deaths within the first month of life. It is generally believed that the etiology of Low Birth Weight is multifactorial. Hence, Low Birth Weight is an important indicator of reproductive health and general health status of population. It continues to remain a major public health problem worldwide especially in the developing countries.

According to World Health Organization (WHO) estimation (2009), it has been reported that about twenty-five million low birth weight babies are born each year, nearly ninety-five percent of them in developing nations. Across the world, neonatal mortality is twenty times more likely for low birth weight babies compared to heavier babies. Then, Low Birth Weight is as a result of preterm birth, intra-uterine growth restriction, or a combination of both pathophysiological conditions. There are numerous factors contributing to Low Birth Weight both maternal and fetal. Low birth weight babies are at higher risk of death (stillborn), disease, miscarriage and disability. The data was extracted from the past records of the obstetrics and gynecology at the

Medical Record Unit of the University College Hospital (UCH), Ibadan, Oyo State over a period of sixty months. It covers two thousand patients whose medical records of delivery were collected by the college.

However, it has been discovered by some scientists that weight at birth is directly influenced by general level of health status of the mother. Thus, maternal environment is the most important determinant of birth weight, and factors that prevent normal circulation across the placenta cause poor nutrient and oxygen supply to the fetus, hence restricting growth. The maternal risk factors are biologically and socially interrelated; most are, however, modifiable.

Therefore, these factors vary from one area to another, depending upon geographic, socio-economic and cultural factors. The mortality of Low Birth Weight can be reduced if the maternal risk factors are detected early and managed by simple techniques. Thus, it is necessary to identify factors prevailing in a particular area responsible for Low Birth Weight. A baby is expected to spend thirty-seven weeks in the womb. An attempt for such baby to come to the world before the expected duration may result into Low Birth Weight situation. With this background in mind, the objective of the present study is to identify the maternal risk factors associated with Low Birth Weight (LBW) in the present world by comparing Classical Least Square (also known Ordinary Least Square) Method of Estimation with Ridge and Principal Components Methods of Regression Analyses.

1.1 Objectives of the Study

The main objectives include:

- To compare estimates of Classical Least Squares (CLS) with that obtained from the application of Ridge and Principal Component Methods of Regression Analyses in multicollinear situation;
- To determine which of the maternal risk factors are responsible for the delivery of low birth weight babies;
- To obtain suitable regression model that best describes multicollinear situation under the present study;
- To determine the degree of linear relationship between low birth weight and the associated factors responsible for it.

1.2 Definitions of Some Terms

Some variables of interest, as used in the present study, require basic simple definitions which are explained below:

* y - **Low Birth Weight:** World Health Organization (WHO) has defined Low Birth Weight as 'one whose birth weight is less than 2.5kg irrespective of the gestational age'. Birth weight is the measurement of the weight of baby immediately after the baby is born.

***** x_1 - **Mother's age:** Age, according to oxford dictionary, is simply defined as the number of years that a person has lived or a thing has existed. Thus, the number of years of existence of a mother at the time of giving birth to newborn baby is referred to as Mother's Age. Nowadays, child-bearing age starts from 13 years; though it might be outlying age of child-bearing as it has been established in some literatures.

***** x_2 - **Mother's weight:** How heavy somebody / something is, which can be measured in, for example, kilograms or pounds is known as weight. Mother's weight is simply the mean weight of the mother (measured in kilogram) before and immediately after giving birth to newborn baby.

***** x_3 - **Mother's height:** Height, as its name implies, is simply the measurement of how tall a person or a thing is. The height of the mother is measured in meter (m) before and after delivery of newborn baby.

***** x_4 - **Preterm delivery:** It is expected of a baby to spend thirty-seven weeks in the womb. If a baby spends less than thirty-seven weeks in the womb before delivery, then such situation is said to be preterm delivery. It is called miscarriage if such baby is delivered as stillborn. Preterm delivery is highly associated with so many factors such as placenta malfunctions; placenta is a nutrient-rich lifeline from mother to infant, so when it is compromised, the infant's growth will drastically suffer. Several types of placenta problems can interfere with a baby's growth. One of them is *placenta previa*, in which the placenta fuses to the cervix, covering all or part of the openings. Even more common is *placenta abruption*, in which the placenta starts separating from the uterine wall during the pregnancy, before delivery. Affecting about two percent of pregnancies, the condition can, in serious cases, reduce the flow of nutrients and oxygen to the baby.

***** x_5 - **Multiple pregnancies:** When a woman carries more than one baby, her risk of premature delivery skyrockets. Each additional baby increases the risk significantly. The preterm rates are approximately 60 percent for twins, 90 percent for triplets, and about 100 percent for quadruplets and beyond (Report from World Health Organization - WHO, 2009). Additional babies stretch the uterus and compete for limited nutrients. Multiple pregnancies also put extra strain on the mother's body, sometimes leading to

complications like anaemia, high blood pressure, and early labour. It's been reported that women between (30-40) years of age are likely to conceive more than one baby at a glance.

* x_6 - **Parity:** The number of babies born by a mother is simply called parity. In this study, records of number of babies of individual women are collected.

* x_7 - **Graphidity:** This is the total number of pregnancies being had by a particular mother; it simply means the addition of stillborns, livebirths and even miscarriages. It is quite different from parity, which is only the number of livebirths alone.

* x_8 - **Maternal infections:** A series of maternal infections are available in the literature. Examples are malaria, tuberculosis, anaemia, intra-uterine infections, congenital abnormalities, etc. It has been reported by some researchers that women who have chronic conditions like diabetes, heart defects, or kidney disease tend to have more difficult pregnancies. As a result, they are more likely to deliver prematurely and have low birth weight babies. Some birth defects can impede normal development of the infants and lead to preterm birth. For example, if an infant develops problems like transposition of the heart's great arteries or *spinal bifida* (open spine) – a condition in which the neural tube fails to close properly- doctors may need to perform surgery while the baby is still in the womb, which raises the risk of preterm birth. A recent study in the Journal of Obstetrics and Gynecology found that the birth defects most commonly associated with preterm delivery include '*Down syndrome*', '*Klinefelter syndrome*', '*Turner syndrome*', '*Patau syndrome*', '*Edwards syndrome*', and congenital structural abnormalities like *orofacial cleft, club foot, polydactyly, hypospadials*. Others include *cardiac, central nervous system* and *musculoskeletal abnormalities.*

* x_9 - **Fetal infections:** For a pregnant mother, catching a common illness like a cold or the flu can raise concerns about harm to the baby. But these everyday illnesses do not typically threaten a developing infant. There are, however, some less common viral and parasitic infections that can indeed cause fetal problems like slow growth, and even birth defects. From the existing literature, we have the followings infections affecting fetus:

(a) Cytomegalovirus: This herpes virus, presents in bodily fluids, is the most common type of virus transmitted to a developing infant. It is associated with disabilities like neural tube defects and Down syndrome.

(b) Rubella: More commonly known as German measles, this virus can cause birth defects like mental retardation and hearing, sight and heart problems. Luckily, German measles can be prevented via the Measles, Mumps and Rubella (MMR) vaccine.

(c) Chickenpox: Exposure to this virus during the first and second trimester is associated with a small chance of congenital varicella syndrome, which can include limb malformation, scarring, growth problems and mental disabilities.

(d) Toxoplasmosis: Infection with this parasite during pregnancy is associated with brain defects and hearing as well as sight loss. The parasite can be present in undercooked meat and cat feces.

II. Methodology

Multiple Linear, Ridge and Principal Component Regressions were used in collaboration with Simple Correlation with a view to comparing results in multicollinear situations. All computations were done with the aid of ***R*** *statistical package*. Data on baby's weight at birth (less than 2.5kg) of approximately two thousand patients were considered as dependent variable while data on mother's age, weight and height, as well as preterm delivery, multiple pregnancies, parity, graphidity, maternal infections such as malaria, tuberculosis, anaemia, intra-uterine infections, congenital abnormalities, etc and fetal infections were used as independent variables.

2.1 Classical Least Squares, Ridge & Principal Component Regression Methods

It has been reported in some literatures that when multicollinearity is present in a set of predictor variables, the Classical Least Squares (also known as Ordinary Least Squares) estimates of the individual regression coefficients tend to be unstable and can lead to erroneous inferences. When this situation arises, several methods have been purportedly designed to handle such scenario. Some researchers believe that data should be transformed using any of the transformation methods such as square transformation, root transformation, reciprocal transformation and so on; while some researchers develop deep interest in other methods of handling multicollinear situations. Principal Component Analysis (PCA), Ridge Regression Approach (RRA), Partial Least Squares Regression (PLSR) to mention but a few are examples of other methods that have been developed to handle situations where explanatory variables are highly correlated. It should be noted that some (if not all) of these methods are biased but still handle such a situation with least mean squares

errors. The least mean squares error property being possessed by these methods makes it shine and better than Ordinary Least Squares estimates. Though, these alternative methods are biased but their estimators tend to have minimum mean squares error than that of the Ordinary Least Squares (OLS) which eventually show efficiency of the results and better precision.

In this research, two alternative estimation methods that provide more informative analyses of the data than the Ordinary Least Squares (OLS) method when multicollinearity is present are considered. The estimators discussed here (ridge estimators) are biased but tend to have more precision (as measured by mean square error) than the OLS estimators (Draper & Smith, 1998; McCallum, 1970; Hoerl & Kennard, 1970). The ridge estimators do not reproduce the estimation data as well as the OLS method; the sum of squared residuals is not as small and, equivalently, the multiple correlation coefficients are not as large. However, any of the alternatives most especially ridge regression and principal component have the potential to produce more precision in the estimated coefficients and smaller prediction errors when the predictions are generated using data other than those used for estimation.

One of the goals of ridge and principal component regressions is to produce a regression model with stable coefficients. The coefficients are stable in the sense that they are not affected by slight variations in the estimation data. Therefore, the main objective of the introduction of ridge and principal component regressions is to select amongst others a set of variables that provides a clear understanding of the process under study as well as formulating an equation that provides accurate forecast of the response variable corresponding to values of the predictor variables.

In Ridge Regression Analysis (RRA), variable selection is done by examining the ridge trace, a plot of the ridge regression coefficients against the ridge parameter/shrinkage estimator, *k.* (Samprit & Hadi, 2006). It has been reported in some literatures that when ***k*=0**, the coefficients in the standardized normal equations are the same with OLS estimates. The parameter *k* may be referred to as the bias parameter. As *k* increases from zero, bias of the estimates increases. As *k* continues to increase without bound, the regression estimates all tend toward zero. The idea of ridge regression is to pick a value of *k* for which the reduction in total variance is not exceeded by the increase in bias. Since *k* is a bias parameter, it is desirable to select the smallest value of *k* for which stability occurs since the size of *k* is directly related to the amount of bias introduced.

It has been reported in the existing literature that there is a positive value of *k* for which the ridge estimates will be stable with respect to small changes in the estimation data (Hoerl and Kennard, 1970). Several methods have been suggested for the choice of the shrinkage estimator *k*. There is a Fixed Point Method where the estimation of shrinkage estimator *k* is suggested to be obtained by the following formular:

$$k = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(0)]^2}$$

where $\hat{\theta}_1(0), \hat{\theta}_2(0), ... \hat{\theta}_p(0)$ are the least squares estimates of $\theta_1, ..., \theta_p$ when the standardized model is fitted to the data (that is, when ***k*=0**), and $\hat{\sigma}^2(0)$ is the corresponding mean square error (Hoerl, Kennard and Baldwin, 1975). Another one is Iterative Method (proposed by Hoerl and Kennard, 1976). In this case, we start with the initial estimate of *k* in the above Fixed Point Method. Denote this by k_0. Then, calculate:

$$k_1 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(k_0)]^2}$$

Then use k_1 to calculate k_2 as follows:

$$k_2 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(k_1)]^2}$$

Repeat this process until the difference between two successive estimates of *k* is negligible. From various researches, it has been discovered that iterative method is preferred to others.

It has been observed that Ridge and Principal Component Regressions are two commonly used biased regression methods. The biased regression methods attack the collinearity problems by computationally suppressing the effect of the collinearity. Ridge Regression does this by reducing the apparent magnitude of the correlations. Principal Component Regression attacks the problem by regressing response variable *Y* on the

important principal components and then parceling out the effect of the principal component variables to the original variables.

III. Analysis of Data

3.1 Verification of Normality Assumption on the Data Sets

Upholding normality assumption is an essential criterion for authenticating analysis of variance in regression analysis. Normal Probability Plot of the Standardized Residual and Histogram of the Standardized Residuals were all used to confirm whether normality assumption is not violated; this is used to confirm the reliability of the result from analysis of variance.

3.1.1 Normal Probability Plot of the Standardized Residuals: This is also called Normal Q-Q Plot. This is the plot of the ordered standardized residuals versus the so-called normal scores. If the residuals are normally distributed, the ordered residuals should be approximately the same as the ordered normal scores. Under normality assumption, this plot should resemble a (nearly) straight line. With the aid of ***R*** statistical package, the plot is drawn and Figure 1 shows that normality assumption is upheld since its appearance is a straight line pattern.

3.1.2 Histogram of the Standardized Residuals: This is another important graph for checking the normality assumption of the residuals. The behaviour of the residuals is an important key in the determination of the normality assumption. The graph is displayed as shown in Figure 2; it is therefore ascertained that the behaviour of the residuals is normal. Hence, normality assumption is not violated.

3.2 Confirmation of Homoscedasticity Assumption on the Data Sets

A situation where the variance of the residuals is affected by at least one predictor variable is simply termed to be heterogeneity/heteroscedasticity. Another important assumption for validity of analysis of variance is equality of error variances. This means that the errors (residuals) must have equal variances. If these errors fail to have equal (sometimes unknown) variances, they will not behave well. Specifically, the assumptions are such that residuals are normally, identically and independently distributed with mean zero and constant but unknown variance leading to the test that all samples came from populations with identical variances (Zar, 1999).

3.2.1 Residual Plot: This is a significant plot for checking homogeneity of sample variances. It has been reported that if the residual plot appears structureless by having about the same extension of scatter of the residuals around zero for each of the variables (under study), it is an indication of homogeneous variances. The plot is drawn as it appears in Figure 3 and it can be deduced from the plot that homogeneity assumption is not violated because of its structureless appearance.

3.2.2 Fligner-Killeen Test: This is newly introduced test for confirming the homogeneity (homoscedasticity) of sample variances. It has been installed in ***R*** statistical package with a view to generating, as appropriate as possible, the test statistic and the p-value. The Fligner-Killeen test has been determined in a simulation study as one of the many tests for homogeneity of variances which is most robust against departures from normality (Conover, Johnson & Johnson, 1981). However, it has been obtained from the package that *p-value* is 2.2 and K-squared of 8.95 with degree of freedom of 8. This is an indication that homogeneity assumption is not violated since *p-value* is greater than value of the type I error (α = 0.05) for which the null hypothesis of presence of homogeneity variances in the data sets is not rejected.

3.3 Test for Existence of Autocorrelation Problem

Autocorrelation problem arises when the errors fail to be independent of each other. It is expected that errors (residuals) should be independent whatever the case may be, but if this condition does not hold as expected, we say there is problem of autocorrelation in the data sets. Therefore, one of the standard assumptions in the regression model is that the error terms are uncorrelated. Correlation in the error term suggests that there is additional information in the data that has not been exploited in the current model. When the observations have a *natural* sequential order, the correlation is referred to as *autocorrelation.*

3.3.1 Durbin-Watson Statistic: This is a test that is aimed at determining whether there is dependency among the successive values of the error term. The most reliable and mostly used test for detecting existence of autocorrelation is Durbin-Watson Test having the statistic:

$$\boldsymbol{d} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_{t-1}^2} = \boldsymbol{1.985} \text{ with } r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_{t-1}^2}$$

An approximate relationship between $\boldsymbol{d}$ and $\hat{r}$ is $\boldsymbol{d} \cong 2(1-\hat{r})$ showing that $\boldsymbol{d}$ has a range of **0** to **4**. Recall that the statistic $\boldsymbol{d}$ is used for testing the null hypothesis $H_0 : r = 0$ against an alternative $H_1 : r > 0$. Note that when $\boldsymbol{r} \cong \mathbf{0}$, the errors are uncorrelated. Thus, from the above value of $\boldsymbol{d}$, we can obtain $\boldsymbol{r}$ as follows:

$$\boldsymbol{d} \cong 2(1-\hat{r})$$

$$1.985 = 2(1- \hat{r}\,)$$

$$\therefore \hat{r} = \mathbf{0.0075} \cong \mathbf{0.01}$$

This can be interpreted that irrespective of the tabular values of Durbin-Watson, we conclude that the presence of autocorrelation in the data sets can be tolerated since the values of *r* and *d* are approximately 0 and 2 respectively.

3.4 Test for Orthogonality of Predictor Variables

Interpretation of the multiple regression model depends implicitly on the assumption that the predictor variables are not strongly interrelated. It is usual to interpret a regression coefficient as measuring the change in the response variable when the corresponding predictor variable is increased by one unit and all other predictor variables are held constant. This interpretation may not be valid if there are strong linear relationships among the predictor variables. It is always conceptually possible to increase the value of one variable in an estimated regression equation while holding the others constant. However, there may be no information about the result of such a manipulation in the estimation data. Thus, it may be impossible to change one variable while holding all others constant in the process being studied. When these conditions exist, simple interpretation of the regression coefficient as a marginal effect is lost. When there is a complete absence of linear relationship among the predictor variables, they are said to be orthogonal.

The seriousness of the effects of multicollinearity seems to depend on the degree of intercorrelation as well as on the overall correlation coefficient. Thus, one might suggest that the standard errors, the partial correlation coefficients and the total R^2 may be used to test for multicollinearity. Yet, none of these criteria by itself is a satisfactory indicator of multicollinearity because of the following reasons:

(i) Large standard errors do not always appear with multicollinearity. (Cobb-Douglas Production Function is very good evidence). Furthermore, large standard errors may arise for various reasons and not only because of the presence of linear relationships among the explanatory variables.

(ii) The intercorrelations of the explanatory variables need not be high for the values of regression coefficients and their standard errors to be affected badly, that is, $r_{x_i x_j}$ is not an adequate criterion by itself.

(iii) The overall R^2 may be high (relative to $r_{x_i x_j}$) and yet the results may be highly imprecise and insignificant (with wrong signs and/or large standard errors).

However, a combination of all these criteria may help the detection of multicollinearity. In order to gain as much knowledge as possible as to the seriousness of multicollinearity, we suggest the application of Farrar-Glauber Test.

3.4.1 Farrar-Glauber Chi-squared Test: A statistical test for multicollinearity has been developed by Farrar and Glauber in one of their books tagged '*Multicollinearity in Regression Analysis'* in 1967. It is really a set of three tests. The first test is *Chi-Square test* for the detection of the existence and severity of multicollinearity in a function. The second test is an *F test* for locating which variables are multicollinear while the third is a *t test* for finding out the pattern of multicollinearity, that is, for determining which variables are responsible for the appearance of multicollinear situations. Only the first test would be considered in this study because we are interested only to know if there exist collinear variables. Farrar-Glauber considered multicollinearity in a sample as a departure of the predictor variables from orthogonality. Therefore, a chi-square test for the presence and severity of multicollinearity in a function with several explanatory variables is outlined as follows:

Hypothesis: H_0 : No multicollinearity versus H_1 : Multicollinearity exists

Test statistic:

$$\chi^2_{cal} = -\left\{(n-1)-\frac{1}{6}(2p+5)\right\}.\ln|R|, \qquad \therefore \chi^2_{cal} = 41.126$$

Conclusion: Since χ^2_{cal} (=41.126) is greater than $\chi^2_{36;0.05}$ (=23.27), we reject H_0. Therefore, it is reasonable to conclude that there is presence of multicollinearity in the data set.

3.5 Estimation by Classical Least Squares Method

The result presented in Table 1 shows that mother's weight and graphidity are the two determinant factors contributing to the delivery of low birth weight babies. But as a result of the absence of orthogonality in the predictor variables, the results of ordinary least squares are imprecise and unreliable. Hence, it is statistically reasonable to suggest that estimates of classical least squares in multicollinear situation are not reliable and cannot be recommended for modelling. Thus, further analysis is suggested to handle such scenario.

3.6 Estimation by Ridge Regression Technique

It has been reported that when multicollinearity is present in a set of predictor variables, the ordinary least square estimates of the individual regression coefficients tend to be unstable and can lead to erroneous inferences. Here, an alternative estimation procedure that provides a more informative analysis of the data than the ordinary least squares method when multicollinearity is present is considered. Ridge Regression provides estimates that are more robust than the least squares estimates for small perturbations in the data. Ridge regression provides a tool for judging the stability of a given body of data for analysis by least squares. In highly collinear situations, as has been pointed out, small changes (perturbations) in the data cause very large changes in the estimated regression coefficients. The method will indicate the sensitivity (or the stability) of the least squares coefficients to small changes in the data.

The ridge estimators are stable in the sense that they are not affected by slight variations in the estimation data. Because of the smaller mean square error property, values of the ridge estimated coefficients are expected to be closer than the OLS estimates to the true values of the regression coefficients. Also, forecasts of the response variable corresponding to values of the predictor variables not included in the estimation set tend to be more accurate. This procedure has been extensively explained in section 2.1 above. Therefore, from ***R*** statistical package, we obtain the following results using Ridge Regression Approach for which the Shrinkage Estimator is 0.004. It is evidenced from Table 3 that only predictor variables (x_2, x_5, x_7, x_8) are significant and adequate for the ridge regression model to be statistically reliable. Therefore, predictor variables with **no** significant explanatory powers have been screened out. The implication of this is that mother's weight, multiple pregnancies, graphidity and maternal infections are the contributing factors responsible for having low birth weight babies with 0.4 per cent biasedness. The value of 0.4 per cent biasedness indicates that the estimates of ridge regression is 99.6 per cent reliable and, are therefore recommended as the best among the three approaches considered. Finally, it is reasonable to model as follows:

$$y = 1.836 + 2.019x_2 + 4.258x_5 + 6.258x_7 + 4.362x_8$$

It's also established, from the same Table 3, that only 18 per cent variations cannot be explained by the estimated model since R^2 = 0.82 as estimated value for coefficient of determination.

3.7 Estimation by Principal Component Method

When predictor variables are considered, the following Correlation Matrix, **R** is obtained as follows:

$$R = \begin{bmatrix} 1 & .992 & .621 & .465 & .979 & .991 & .971 & .674 & .657 \\ & 1 & .604 & .446 & .991 & .995 & .984 & .874 & .597 \\ & & 1 & -.177 & .687 & .664 & .816 & .673 & .832 \\ & & & 1 & .364 & .417 & .457 & .765 & .976 \\ & & & & 1 & .995 & .839 & .518 & .692 \\ & & & & & 1 & .971 & .754 & .682 \\ & & & & & & 1 & -.582 & .785 \\ & & & & & & & 1 & .652 \\ & & & & & & & & 1 \end{bmatrix}$$

From the correlation matrix, R, we obtained the following eigenvalues and the corresponding eigenvectors as follows:

$$\lambda_1 = 6.628, \quad V_1 = \begin{bmatrix} 0.378 \\ 0.383 \\ 0.295 \\ 0.225 \\ 0.366 \\ 0.384 \\ 0.334 \\ 0.260 \\ 0.336 \end{bmatrix}, \quad \lambda_2 = 1.757, \quad V_2 = \begin{bmatrix} 0.074 \\ 0.013 \\ 0.155 \\ -0.443 \\ 0.119 \\ 0.058 \\ 0.546 \\ -0.655 \\ -0.169 \end{bmatrix}, \quad \lambda_3 = 1.156, \quad V_3 = \begin{bmatrix} -0.044 \\ -0.128 \\ -0.426 \\ 0.648 \\ -0.106 \\ -0.108 \\ 0.342 \\ -0.370 \\ 0.322 \end{bmatrix}$$

$$\lambda_4 = -1.035, \quad V_4 = \begin{bmatrix} 0.134 \\ 0.265 \\ 0.290 \\ 0.258 \\ -0.025 \\ 0.178 \\ -0.616 \\ -0.591 \\ 0.015 \end{bmatrix}, \quad \lambda_5 = 0.782, \quad V_5 = \begin{bmatrix} -0.270 \\ -0.332 \\ 0.648 \\ -0.038 \\ -0.191 \\ -0.220 \\ 0.051 \\ -0.008 \\ 0.554 \end{bmatrix} \quad \lambda_6 = -0.366, \quad V_6 = \begin{bmatrix} -0.106 \\ -0.218 \\ 0.432 \\ 0.513 \\ 0.112 \\ -0.066 \\ 0.192 \\ 0.081 \\ -0.657 \end{bmatrix}$$

$$\lambda_7 = 0.066, \quad V_7 = \begin{bmatrix} 0.121 \\ 0.407 \\ 0.114 \\ 0.030 \\ 0.854 \\ 0.039 \\ 0.222 \\ 0.094 \\ 0.128 \end{bmatrix}, \quad \lambda_8 = 0.009 \quad V_8 = \begin{bmatrix} 0.856 \\ -0.415 \\ 0.030 \\ -0.021 \\ -0.103 \\ -0.279 \\ -0.072 \\ 0.008 \\ -0.007 \end{bmatrix}, \quad \lambda_9 = 0.002, \quad V_9 = \begin{bmatrix} -0.012 \\ -0.521 \\ -0.055 \\ -0.022 \\ -0.225 \\ 0.821 \\ -0.021 \\ 0.006 \\ -0.005 \end{bmatrix}$$

Therefore, the following Principal Components are formed:

$$\begin{aligned}
P_1 &= 0.378Z_1 + 0.383Z_2 + 0.295Z_3 + 0.225Z_4 + 0.366Z_5 + 0.384Z_6 + 0.334Z_7 + 0.260Z_8 + 0.336Z_9 \\
P_2 &= 0.074Z_1 + 0.013Z_2 + 0.155Z_3 - 0.443Z_4 + 0.119Z_5 + 0.058Z_6 + 0.546Z_7 - 0.655Z_8 - 0.169Z_9 \\
P_3 &= -0.044Z_1 - 0.128Z_2 - 0.426Z_3 + 0.648Z_4 - 0.106Z_5 - 0.108Z_6 + 0.342Z_7 - 0.370Z_8 + 0.322Z_9 \\
P_4 &= 0.134Z_1 + 0.265Z_2 + 0.290Z_3 + 0.258Z_4 - 0.025Z_5 + 0.178Z_6 - 0.616Z_7 - 0.591Z_8 + 0.015Z_9 \\
P_5 &= -0.270Z_1 - 0.332Z_2 + 0.648Z_3 - 0.038Z_4 - 0.191Z_5 - 0.220Z_6 + 0.051Z_7 - 0.008Z_8 + 0.554Z_9 \\
P_6 &= -0.106Z_1 - 0.218Z_2 + 0.432Z_3 + 0.513Z_4 + 0.112Z_5 - 0.066Z_6 + 0.192Z_7 + 0.081Z_8 - 0.657Z_9 \\
P_7 &= 0.121Z_1 + 0.407Z_2 + 0.114Z_3 + 0.030Z_4 + 0.854Z_5 + 0.039Z_6 + 0.222Z_7 + 0.094Z_8 + 0.128Z_9 \\
P_8 &= 0.856Z_1 - 0.415Z_2 + 0.030Z_3 - 0.021Z_4 - 0.103Z_5 - 0.279Z_6 - 0.072Z_7 - 0.008Z_8 - 0.007Z_9 \\
P_9 &= -0.012Z_1 - 0.521Z_2 - 0.055Z_3 - 0.022Z_4 - 0.225Z_5 + 0.821Z_6 - 0.021Z_7 + 0.006Z_8 - 0.005Z_9
\end{aligned}$$

We shall calculate the Percentage Contributions of each of the Principal Components as follows:

$$P_i = \left[\frac{Var(P_i)}{\sum_{i=1}^{9} \lambda_i} .100 \right] \% \quad for \quad which \sum_{i=1}^{9} \lambda_i = 8.999$$

$P_1 = 73.66\%$ ($with\ \lambda_1 = 6.628$), $P_2 = 19.52\%$ ($with\ \lambda_2 = 1.757$), $P_3 = 12.85\%$ ($with\ \lambda_3 = 1.156$),
$P_4 = -11.50\%$ ($with\ \lambda_4 = -1.035$), $P_5 = 8.69\%$ ($with\ \lambda_5 = 0.782$), $P_6 = -4.07\%$ ($with\ \lambda_6 = -0.366$),
$P_7 = 0.73\%$ ($with\ \lambda_7 = 0.066$), $P_8 = 0.10\%$ ($with\ \lambda_8 = 0.009$), $P_9 = 0.02\%$ ($with\ \lambda_9 = 0.002$).

3.8 Determination of the Number of Principal Components to be included in the Analysis

Several methods are available in the literature as criteria for the determination of number of components to be extracted. Two of the methods are discussed one after the other as follows:

3.8.1 Kaiser's Criterion: This decision rule has been suggested by Guttman and adapted by Kaiser. Its application is simple; only Principal Components having latent roots (eigenvalues) greater than one (1) are considered essential and should be retained in the analysis. In other words, we retain P_i if $\lambda_i \geq 1$. Therefore, the first three components are essential and significant by Kaiser's Criterion.

3.8.2 Cattell's Screen Criterion: In this case, we plot latent roots (eigenvalues) against the order of extraction of the *P's* to be retained in the analysis. Therefore, we use the shape of the resulting curve to judge how many *P's* to retain. The decision rule is to retain the *P's* up to the point where the resulting curve has some curvature and reject the *P's* for which the curve becomes a straight line. In other words, the point at which the curve

straightens out (develops into a linear relationship between the order of *P's* and their latent roots/eigenvalues) determines the maximum number of *P's* to be extracted. Beyond this point, the *P's* are unreliable because they are heavily affected by factors which are not common to all *X's.* It is evidenced from the plot shown in Figure 4 that only the first three components are significant and should be retained.

3.9 Obtaining Principal Component Regression (PCR)

From ***R*** statistical package, the following estimates have been obtained after which each explanatory variable has been standardized in each of the three components. Therefore, from Table 4, it is inferred that only P_3 is significant and the final principal component regression is obtained as follows: $y = 2.503 - 0.406P_3$.

IV. Conclusion

Factors associated with low birth weight, often termed as risk factors and their presence in an individual woman indicates an increased chance or risk of bearing a low birth weight infant. Globally, low birth weight, as an indicator, is a good summary measure of a multifaceted public health problem that includes long-term maternal malnutrition, ill-health, hard work and poor pregnancy health cares. Besides these, it is also associated with multi-pregnancy, genetic history (previous preterm birth and/or abortion), smoking or exposure to second hand smoking, stress or lack of support, maternal infections such as bladder or virginal infections which can cause labour to start early, stressful work condition such as being on feet for long hours, self drug administration and drinking alcohol as well as smoking cigarette, being under-weight before pregnancy, poverty, mother's age and parity to mention but a few.

When there is a complete absence of linear relationship among a set of predictor variables, then they are said to be orthogonal. The check for the existence of multicollinearity problem among the predictor variables is not left-out. From the application of Farrar-Glauber Test, the result obtained shows that multicollinearity is suspected in the data sets. Also, the degrees of relationships among some predictor variables are very high invariably meaning that some explanatory variables are highly collinear. From Table 3, there is a very high positive relationship of about 99% between mother's age and parity; likewise between mother's weight and multiple pregnancies. It's between mother's height and preterm delivery that lowest positive degree of relationship among the response variables was recorded.

With two thousand cases examined, results obtained from the application of Classical Least Squares on the dataset would have been the best if not the presence of multicollinearity problem. The model obtained by stepwise selection through the application of ordinary least square would have been the best if not non-orthogonality of the predictor variables. The value of shrinkage estimator was computed to be 0.004 meaning that we are 0.4% biased with our estimation when Ridge Method of Regression is applied. However, results obtained from the application of Ridge Method of Regression (Table 3) show that only four predictor variables are statistically significant with approximately 82% variation in low birth weight is explained by mother's weight, multiple pregnancies, graphidity and maternal infections like malaria, tuberculosis, anaemia/shortage of blood, intra-uterine infections, congenital abnormalities, etc.

However, from the formation of principal components, it's deduced that out of nine components formed; only the first three are extracted. The first component has about 73.66% positive contribution while two of the components contributed negatively. From Kaiser's and Cattell's Screen Criteria, it is has been established that the first three components contributed significant and are chosen to be included in the analyses. The results obtained from obtaining Principal Component Regression revealed that only the third component is statistically significant. However, note that one cannot easily assign any specific economic meaning to the new variable in principal components. It is an artificial orthogonal variable not directly identifiable with a particular meaning; this is because all predictor variables are embedded in the new variable. It has been established that the method uses less of the information contained in the sample since the number of *P's* retained is far smaller than the number of predictor variables. This is the reason why some researchers consider this method of principal component inappropriate as a solution to multicollinearity problem but manageable. The *P's* are artificial variables to which no specific meaning can be assigned. As known that the method of principal component can be applied either on the original *X's*, or on their deviations from means, or on the standardized values of the original variables, therefore the values of the principal components will be different in each case. This dependence on the unit of measurement is obviously a serious weakness of the Principal Component Techniques.

In a nutshell, Ridge Regression Method is preferred to be the best method of parameter estimation in a multicollinear situation provided that the amount/percentage of biasedness is bearable. In this study, it is obvious that 0.4% of biasedness can be tolerated. Therefore, the suitable regression model that best describes the situation of multicollinearity is: $y = 1.837 + 2.019x_2 + 4.258x_5 + 6.258x_7 + 4.362x_8$. In summary, the major factors responsible for having low birth weight babies are weight of the mother, having more than one

pregnancy at a time (multiple pregnancies), total number of pregnancies already had (graphidity) and associated maternal infections like malaria, tuberculosis, anaemia, intra-uterine infections, congenital abnormalities, and so on.

Recommendations

In order to reduce (if not completely eliminate) the rate of having low birth weight babies among our women, the following recommendations are made:

(a) It is advisable for women to eat balanced diet with a view to gaining more weight during pregnancy; they should consume more fruits and vegetables and, less of the junk foods;
(b) There should be no drinking of alcohol, self-drug administration and no smoking before, during and even after pregnancy;
(c) Women are advised to watch their weights, undertake regular low impact exercise and take time to relax;
(d) Regularly consulting doctors/nurses in charge of ante-natal can still be used as a measure to curb having low birth weight babies during pregnancy;
(e) Hence, many risks for low birth weight can be identified before pregnancy occurs. Thus, health education, socio-economic development, maternal nutrition, and regular use of health services during pregnancy are all important for reducing low birth weight;
(f) Reduction in the rate of abortions among youths of adolescent age and the rate of child-bearing can be used to reduce or even eliminate having low birth weight babies.

References

[1]. Berghella V. Prevention of Recurrent Fetal Growth Restriction. Obs and Gyn 2007; 110(4): 904-112
[2]. Brown P. J. (1994), Measurement, Regression and Calibration, Oxford
[3]. Conover W. J., Johnson M. E. and Johnson M. M. (1981): A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, Technometrics (23): 351-361
[4]. Deswal B. S., Singh J. V., Kumar D.- A Study of Risk Factors for Low Birth Weight, Indian J. Community Med. 2008; 24: 127-131
[5]. Development Core Team (2012): R: A Language and Environment for Statistical Computing, Vienna, R Foundation for Statistical Computing
[6]. Draper N. R., Smith H.- Applied Regression Analysis. Wiley Series in Probability and Statistics, 1998
[7]. Martin J. A.- Births: Final Data for 2005. National Vital Statistics Reports; 2007 Dec. Report No. 6(56)
[8]. Oxford Advanced Learner's Dictionary of Current English by A. S. Hornby, seventh edition
[9]. Smith G. C., Pell J. P., Debbie R.: Caesarean section and risk of unexplained stillbirth in subsequent pregnancy. Lanat 2010, 362(9398): 1779-1784
[10]. UNICEF and WHO, Low Birth Weight; Country, Regional and Global Estimates, New York, WHO Bulletin, 2009, 83(3): 178-185
[11]. Zar, J. H. (1999): Biostatistical Analysis, Prentice Hall, New Jersey, 4th ed.

Authors' Biography

Olawuwo Simeon[1]: A seasoned teacher and researcher. He holds Bachelor of Science in Mathematics Education (1991) from University of Ilorin, Nigeria; Master of Science Degree (2002) in Demography and Social Statistics from Obafemi Awolowo University, Ile-Ife, Nigeria and Master of Science Degree (2012) in Biostatistics from University of Ibadan, Nigeria. He is a Principal Lecturer in the Department of Statistics, Federal Polytechnic Ede, Nigeria. He has taught many statistical and mathematical courses at various higher institutions of learning. He can be contacted through this mail: soolawuwo@yahoo.com.

Ogunleye Timothy A.[2]: He was born in the city of Ibadan, Nigeria some years ago. He hails from Ogbaagbaa Township in Ola-Oluwa Local Government Area of the State of Osun, and currently working in one of the local governments in Nigeria. He holds Higher National Diploma certificate in Statistics from the Federal Polytechnic Ede, Osun State in 2004. He later obtained Bachelor of Science [B.Sc. (Hons)- 2nd Class Upper Division] and Master of Science Degrees in Statistics both from University of Ilorin, Nigeria in the years 2011 and 2014 respectively. His research areas are Modelling, Econometrics, Multivariate Analysis and Experimental Design. Having been graded with over sixty percent in the M.Sc. Examinations, he is interested to proceed with Ph.D. programme soonest. His electronic mail is thompsondx@gmail.com.

Ojo Thompson O.[3]: A seasoned teacher and researcher. He holds Bachelor of Science (Hons) (1995) and Masters of Science (2001) Degrees in Statistics from University of Ibadan, Nigeria and Doctor of Philosophy in Statistics from University of Botswana (2013). He is a Senior Lecturer in the Department of Statistics, Federal Polytechnic Ede, Nigeria. He has taught many statistical courses at various higher institutions of learning. He can be contacted through the email: bodtommy2005@yahoo.com.

Adejumo Adebowale O.[4]: A senior lecturer in the university, who holds B.Sc. (Hons) Statistics (2nd Class Upper Division) (1995), Ilorin; M.Sc. Statistics (1998), Ilorin; Ph.D. Statistics (Dr. rer. nat), Ludwig-

Maximilian-University (LMU), Munich, Germany (2005). His research interests are Modelling, Time Series, Biostatistics and Categorical Data Analysis. His electronic mail is aodejumo@gmail.com.

Acknowledgement:

The authors wish to express sincere appreciation to the management and staff of University Teaching Hospital (UCH), Ibadan, Nigeria for making available relevant data for this research work. We are particularly grateful to Dr. Isamotu Rafiu-Gazette (Special Assistant to Osun State Governor on Health Matters), Prof. Babatunde Lateef Adeleke (Dean, Faculty of Science, University of Ilorin, Nigeria), Pastor R. M. Alabison (Dean, School of Pure and Applied Sciences, Federal Polytechnic Ede, Nigeria), Pastor Afolabi Samuel Olufemi (Head, Accountancy Department, Federal Polytechnic Ede, Nigeria) and Dr. Kolapo Ige (Head, Department of Economics, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria) for all efforts put forward to achieve success. We are also grateful to Dr Alfred A. Abiodun, Dr Gafar M. Oyeyemi (both from Department of Statistics, University of Ilorin, Nigeria) and Elder Adewusi Adegboyega Adelakin (Director of Budget, Planning, Research and Statistics Department, Ejigbo Local Government, Osun State, Nigeria) for all inputs. We are grateful to Pa Paul Adeboye Ogunjimi, Engr. Hassan Oladepo Adedapo (Head of Local Government Administration- Ejigbo LG), Mrs Adeola O. Olafare (Director of Administration and General Services, Ejigbo LG), Mrs Oyedeji Folasade (Finance and Supplies Department in Ifelodun LG), Mr Salotun Augustus Adebayo (Administrative Officer, Federal Polytechnic Ede, Nigeria), Adegbenro Abraham Bukola of Success Business Venture, Oja-Timi, Ede, State of Osun and my educational benefactor Oyeterun Oluwafemi Samuel. We wish to appreciate Mrs Adebodun Racheal O. (Department of Budget, Planning, Research and Statistics, Ede North Local Government, Nigeria), Ogunleye Samson Oluseyi, Dr Wahab Babatunde Yahya (Head of Department of Statistics, University of Ilorin, Nigeria), Dr Amos A. Adewara, Prof. Emmanuel Teju Jolayemi, Prof. Peter A. Osanaiye and Prof. Ben. A. Oyejola for different kinds of contribution made on the success of this journal. Thank you all and God bless!

APPENDIX

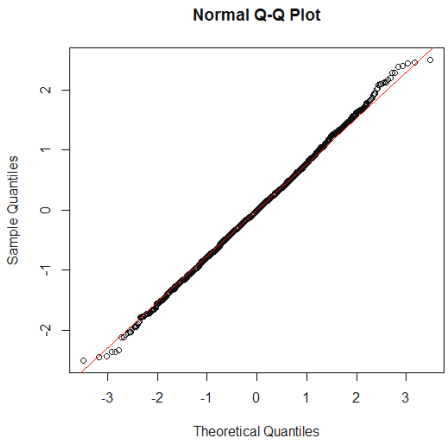

Figure 1: Normal Q-Q Plot

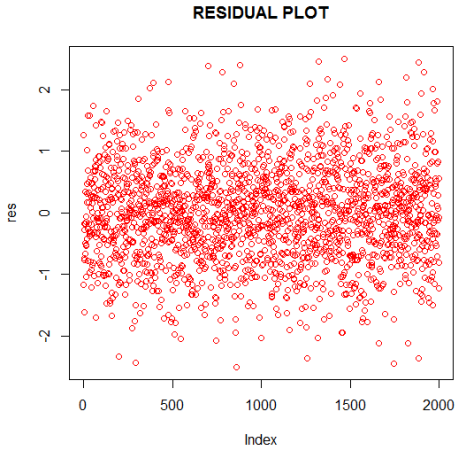

Figure 3: Residual Plot

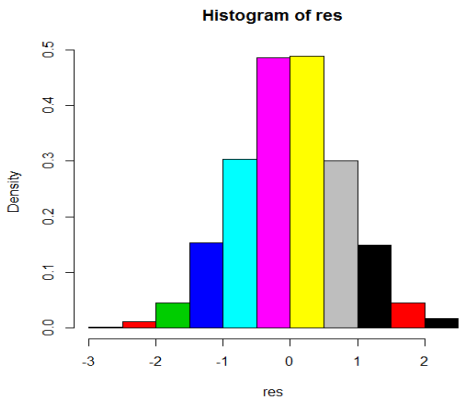

Figure 2: Histogram of the Standardized Residuals

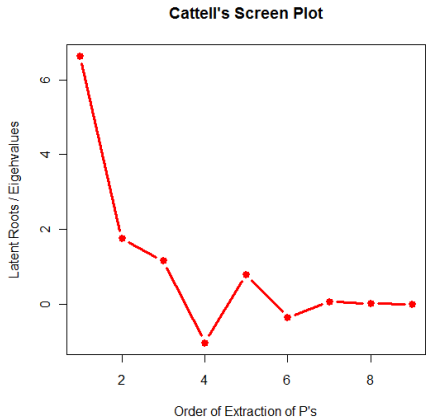

Figure 4: Cattell's Screen Plot

TABLE 1: Upper Triangular Correlation Matrix Using OLS

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Y
x_1	1	.992	.621	.465	.979	.991	.971	.674	.657	.375
x_2		1	.604	.446	.991	.995	.984	.874	.597	.846
x_3			1	-.177	.687	.664	.816	.673	.832	.826
x_4				1	.364	.417	.457	.765	.976	.219
x_5					1	.995	.839	.518	.692	.658
x_6						1	.971	.754	.682	.769
x_7							1	-.582	.785	.811
x_8								1	.652	.556
x_9									1	0.004
Y										1

TABLE 2: Parameters of Least Squares Results (Full Model)

Model	Coefficients	Std. Errors	P-Values
Intercept	1.976	22.474	0.930
x_1	0.041	0.034	2.226
x_2	-0.018	0.374	0.013
x_3	0.501	0.234	3.033
x_4	-0.029	0.050	2.557
x_5	-0.053	0.059	2.037
x_6	-0.019	0.041	1.944
x_7	-0.085	0.040	0.034
x_8	0.064	0.051	3.528
x_9	-0.030	0.036	2.401

TABLE 3: Ridge Regression Parameters

Model	Coefficients	Std. Errors	P-Values
Intercept	1.836	0.474	0.001
x_1	0.329	21.034	0.818
x_2	2.019	0.068	0.000
x_3	4.032	8.234	0.891
x_4	-8.291	24.958	0.861
x_5	4.258	0.824	0.001
x_6	-1.217	17.918	0.721
x_7	6.258	0.718	0.010

x_8	4.362	0.293	0.009
x_9	-9.932	61.902	0.994
R^2 = 0.82, with 2000 cases			

TABLE 4: Parameters of Principal Component Regression

Model	**Coefficient**	**Std. Error**	**t-value**	**P-Value**
Intercept	2.503	0.018	141.114	*2e-16****
P_1	-0.003	0.018	-0.146	0.884
P_2	-0.022	0.017	-1.269	0.204
P_3	-0.046	0.018	-2.562	0.010***
R^2 = 0.62 with 2000 cases				