Predictor Envelopes and Standard Regression Models: An Empirical Juxtaposition

Joseph Bamidele Odeyemi

Department of Mathematics/Statistics, Federal Polytechnic, Offa, Nigeria

Abstract: ***This study compares predictor envelopes and standard regression models using data on the attributes (Weight, Height, Shoe Size, Chest Diameter, Diastolic Blood Pressure, Fasting Blood Sugar, and Age) of pregnant women obtained from General Hospital, Offa. The purpose is to use both methods to fit models and determine which of the two is more efficient in prediction. The study applied predictor envelopes method as well as standard regression method to the data and was analyzed with the support of Renvlp statistical software. It was found that the information criterion AIC and BIC agreed to u=2 sufficient dimension. It was also discovered that both methods declare variable height as the only active predictor of weight of pregnant women due to their estimates (51.07, 52.03), z-scores (0.44, 0.45), and ratios of asymptotic standard error (0.99) The predictive performance of the two models show that the standard error of predictor envelopes estimates are smaller than those of the standard regression estimates. Thus, Predictor envelope is better than standard regression method even with one response variable y.***

Keywords: Renvlp software, standard regression, sufficient dimension, predictor envelopes

1. Introduction

Generally, hospitals are expected to monitor the number of diseases that occur in its different patients care units every time. Infections can adversely affect treatment and can sometimes result to death. Statistical modeling has been used to monitor the occurrence of diseases in the past. Akpa and Oyejola (2006) used statistical modeling to model HIV/AIDS epidemics. Also, Adeyemi (2007) applied statistical monitoring model to medical administration statistical model is a powerful tool in healthcare services.

Several methods have been developed for statistical modeling of a set of data which involve a set of procedures for estimating the relationships among variables. One of these methods is regression. The purpose of statistical regression modeling is prediction, or error reduction, or to explain variation in response variable attributed to variation in explanatory variables. Regression is used to fit a predictive model to an observed data set of values of the response in order to determine the predictors that are most important for determining the response variable. But it is usually difficult to determine which of the predictors is most important in determining the value of the response variable. One basic solution is to estimate the model using different methods which often produce several models.

In the words of Konishi and Kitagawa (2008) it was affirmed that problem of estimation is related to the problem of statistical modeling typically formulated for the purpose of comparisons. Models are compared to each other by exploratory data analysis. This is a situation where different types of models are formulated and an assessment is performed of how well each one describes the data. Some common criteria for comparing models include R2, Bayes factor, AIC, BIC, and likelihood ratio test. Each of these will help us to make statements about the elements of the predictor which is likely to adequately approximate the true distribution since there is a probability distribution underlying the observed data induced by the process that generated the data.

After fitting a number of regression models to a given data set, we may measure the errors involve in each of the estimated model. A measure of error in statistics is the root mean squared error (RMSE) which is the root of mean squared error. When this is adjusted for the degree of freedom for error it is called the standard error of estimate in regression analysis. It is this value that we minimize in estimation process because it is this value that determines the width of the confidence intervals for predictions. The root mean squared error that is smaller is used to judge between two models. Other error measures used to compare the performance of models in absolute or relative terms are the mean error (ME), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). The mean error is a signed measure of error which indicates whether the predictions are biased. It is a component of mean squared error. In a model that includes constant term mean squared error is minimized when mean error equals zero. Another known criterion on which model comparisons can be based is to carry out residual diagnostics and goodness of fit tests. But experience had shown that a model should not be chosen over another on these bases. Other criteria include simplicity and usefulness for decision making of the model. However, it is sometimes hard to know which comparisons are most important. In model selection and regularization problem, dimension reduction methods have been used as alternative procedures to using least squares to fit linear model and found to be better than methods which use subset of original variables or those who shrink their coefficients toward zero. This is to enhance the predictive and interpretability of the model. By dimension we mean the maximum number of independent vectors in a set of non empty set vectors in a matrix of data. Dimension reduction methods involve combining a number of predictors into a certain number of dimensional subspaces, thus reducing the number of coefficients to be estimated. These computed combinations of predictors are then used as predictors to fit a linear regression model by least squares. In other words we fit least squares model after the predictors have been transformed. Every dimension reduction method is reported to be using these two steps. Some of these methods are the principal component regression, partial least squares, to mention but two. However, it has also been asserted that if the number of variables is equal or more than the observations least squares usually report perfect fit regardless of whether there is a relationship or not. Therefore, its use in such a situation is not permitted.

Su and Lee (2018) observed that in many regressions the response is one dimensional and the number of predictors is usually more which may be in multiples and many (large). In the context of sufficient dimension reduction, these regressions are usually viewed from two angles. Firstly, full regressions, where all the predictors are said to be carrying information about the response, and sparse regressions, where only some of the predictors are said to be carrying information about the response. Sufficient dimension reduction, in line with this, attempts to identify the smallest possible number of linear combinations of predictors known as sufficient predictors which retain all the necessary information in the predictors about the response distribution. Literatures have shown that initially, many of the earlier dimension reduction methods were not based on sufficiency but as the field of statistics grew dimension reduction methods also grow with it. Sufficiency was later incorporated into dimension reduction methods. This involves the use of fewer linear combinations of predictors to achieve dimension reduction without loss of information. Sufficient dimension reduction involves new ideas and techniques of reducing dimensions. The method has a lot of applications in regressions. We can use it in regression graphics to find a sufficient summary plot that gives all necessary regression information. Sufficient dimension reduction is used to extracts information in regressions in which the number of feature is less, equal or greater than the number of observations in order to predict or classify response variables. In sufficient dimension reduction some predictors can be left out without losing any information which we can achieve through conditional independence. Cook (2018) has observed that sufficient dimension reduction organizes the variations of data in a predictable and interpretable way like some other reduction methods and it has a unique feature such that the method organizes these variations existing in the predictors according to how much they can explain the response distribution. Similarly, sufficient dimension reduction is known to look for only the variables that can predict the response well, thus reducing the number of predictors for the response. It is then aimed at achieving low rank. Sufficient predictors may be non linear functions or linear, continuous or categorical, vector or matrix. The responses can both be functions or vectors of functions as well. The basis of reduction may be means, quantiles or conditional variances.

The envelope model proposed by Li (2018) in the field of sufficient dimension reduction is popularly used to identify the central subspace consistently. It has the potential to gain efficiency in estimation. In the time past several investigations have been done using different methods of parameter estimation. In spite of this it is not certain if any of the methods have been proved to negate the potency of envelope method, because they have not exhaustively estimated the entire central subspace than the envelope model. Dimension in this work refers to the maximum number of independent vectors in a set of non empty set vectors in a matrix of data. Dimension reduction also refers to all methods used in combining a number of predictors into a certain number of dimensional subspaces. Thus, reducing the number of coefficients to be estimated (Li, 2018). These computed linear combinations of predictors are used as predictors to fit a linear regression model by least squares. In other words, we fit least squares model after the predictors have been transformed. Majority of dimension reduction methods involves linear combinations of the predictors which are projection of the predictors onto subspaces. These subspaces are referred to as dimension reduction subspaces. Under some conditions the intersection of all these dimension reduction subspaces is referred to as sufficient dimension (Cook 1998). This is a way of identifying p-vector predictor that largely has information connecting with a response in a matrix of data. This is done through the evaluation of all the p-vector predictors in the matrix and separating those containing immaterial information and accounting for it in the one with material information.

2. Problem Statement

Statistical models are usually part of the foundation of inferences made about a given data set. Many statistical hypothesis tests and statistical estimators are derived from models. There have been many efforts put in place by statisticians for statistical modeling. These include monitoring models, model evaluation, model comparison, model selection regularization, model interpretation, model building techniques, and many others. These not withstanding, we can not say we have achieved the best in modeling because researchers are still finding difficulties in choosing methods for building, using and interpreting suitable models. Many researchers use models wrongly. T he fact is that when variables for building a model are many researchers are disturbed. Many questions come to their heart. These questions need answers. What statistical model could one use to measure the impact of personality traits on a single outcome variable with only one level? What is the best approach to select input variable in building model? How can one calculate effect of predictor variable from a linear model? What is the statistical relationship between response and explanatory variables? What are some useful methods to statistically compare models? What is the common practice in model comparison? How well does my model fit my data? Is my model adequate? Why is the prediction from my model not reliable? What is the statistical model I use for my data analysis? Therefore, model identification and evaluation are inevitable.

3. Objectives of the Study

This study seeks to compare envelopes and standard regression models using hospital data. The specific objectives are:

1) To demonstrate use and superiority of predictor envelope models to regression
2) To show efficient coefficient estimation of predictor envelopes compared to standard regression model

4. Literature Review

Li, Bondell, and Reich (2010) gave their report on sufficient dimension reduction using Bayesian mixture modeling that the technique of Bayesian was efficient and provides a unified framework to handle categorical predictors, missing predictors and Bayesian variable selection. Yin and Hilalu (2015), worked on sequential sufficient dimension reduction for large P, small n problems in which they proposed a new and simple framework for dimension reduction in the large p, small n setting. The framework permits the existing approaches for n>p to be adapted to n<p problems. They suggested a sufficient procedure path using sufficient dimension reduction techniques which to them is very general.

Zhang and Cook (2014) in their approach to sufficient dimension reduction extended envelope method proposed for reducing estimative and predictive variations in regression by Chiaromonte. They proposed a general definition of an envelope as well as a general framework for adapting envelope methods to any estimation procedure. Their framework was demonstrated on weighted least squares, generalized linear models and Cox regression. They demonstrated the potential of envelope methods to improve standard methods in logistic regression, poisson regression and linear discriminant analysis. Su and Lee (2018) demonstrated the use of R package for efficient estimation in multivariate models and concluded that R package captured all the standard methods in any form for both model free and model based dimension reduction and improve on them. It was discovered from their work that R software that is based on envelope method provides model fitting and inference functions for bootstrapping, cross validation, prediction and hypothesis testing.

5. Methodology

i) Model specification

The standard regression models where Y and X are said to be jointly distributed according to Cook (2018) could be any of the following:

(a) Linear regression:- $Y = \alpha + \beta X + \varepsilon$

(b) Nonlinear regression:- $Y = \alpha + \alpha_1 \ell^{-\beta X + \varepsilon}$

(c) Logistic regression:- $\log it(P) = \alpha + \beta X$

(d) Nonlinear logistic regression:- $\log it(P) = \alpha + f(\beta_1 X + \beta_2 X)$

In Cook (2018) envelopes are reported to be specialized forms of sufficient dimension reduction methods. They are said to be applicable in model based analyses. The formal definition of an envelope as put by cook goes thus:

Let $M \in S^{r \times r}$ and let $S \subseteq span(M)$. Span in this sense means linear combinations of vectors of the same components. A span is a vector space. The M-envelope of S, written as $\varepsilon_M(s)$, is the intersection of reducing subspaces of M that contain S. This is the intersection of all the linear combinations of vectors which contains S.The envelope model proposed by Cook, Li, and Chiromonte (2010) is used to identify the immaterial information. The predictor envelope model in Cook, et al (2013) is used for dimension reduction of predictors in the context of linear regression

$$Y = \mu + \beta X + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where $Y \in \Re^r$ is the response vector, and $X \in \Re^p$ is the stochastic predictor vector with mean μ_x, and Σ_x. The error vector ε has mean 0 and covariance matrix $\Sigma_{y/x}$. The regression coefficient is contained in $\beta \in \Re^{pxr}$ and $\mu \in \Re^r$ is the intercept. Y can be univariate or multivariate response vector. Following the convention in Cook et al, 2013 the earlier model is reformulated to describe predictor envelope model as $Y = \mu + \beta^T(X - \mu_x) + \varepsilon$.

Following the convention in Cook et al. 2013, model (1) is reformulated to describe predictor envelope model as

$$Y = \mu + \beta^T(X - \mu_X) + \epsilon, (2)$$

Let S denote a subspace of $\mathbb{R}^p$. Predictor vector X is decomposed into a material part $P_S X$ (X-variant) and an immaterial part $Q_S X$ (X-invariant) such that the following two conditions are satisfied:
$COV(Y, Q_S X|P_S X) = 0$;and$COV(P_S X, Q_S X) = 0$.

These two conditions imply that $Q_S X$ does not affect the distribution of Y directly or through the association with $P_S X$.

The two conditions are equivalent to $(i)' span(\beta) \subseteq S$ and $(ii)'' \sum_X = P_S \sum_X P_S + Q_S \sum_X Q_S$.And the intersection all S that satisfies $(i)'$ and $(i)''$ is the $\sum_X$-envelope of β, denoted by $\mathcal{E}_{\sum_X}(\beta)$, or simply $\mathcal{E}$.

Its dimension is denoted by u.Let$\Gamma \in \mathbb{R}^{p \times u}$ be an orthonormal basis of $\mathcal{E}_{\sum_X}(\beta)$, and $\Gamma_0 \in \mathbb{R}^{p \times (p-u)}$ be its completion. Under conditions $(i)'$ and $(i)''$, (2) can be reformulated as

$$\mathrm{Y} = \mu + \eta^{\mathrm{T}} \Omega^{-1} \Gamma^{\mathrm{T}} (\mathrm{X} - \mu \mathrm{X}) + \epsilon, \sum_{\mathrm{X}}$$
$$= \Gamma \Omega \Gamma^{\mathrm{T}} + \Gamma_0 \Omega_0 \Gamma_0^{\mathrm{T}}, (3)$$

Where $\beta = \Gamma \Omega^{-1} \eta, \eta \in \mathbb{R}^{u \times r}$, and $\Omega^{-1} \eta$ carries the coordinates of β with respect to Γ.Thematrice$\Omega \in \mathbb{R}^{u \times u}$ and $\Omega_0 \in \mathbb{R}^{(p-u) \times (p-u)}$ are positive definite. When $u = p$, the predictor envelope model (3) reduces to the linear regression model (1) $u = 0, \beta = 0$ and the covariance between X and Y is zero. Determination of number of dimension u is by information criteria

ii) Data Source and Summary

The data set utilized in this study was extracted from records of general hospital, Offa, in Kwara State of Nigeria. The method of collection is transcription method. There are 150 observations for each variable. The variables represent the personal traits of pregnant women patients visiting the hospital. The data come from the routine activities of the medical personnel of this hospital.

The data comprises information on seven variables as summarized in the table below.

Table 1: Data Summary

Variable	Description
Response (Y)	Weight (kg)
X_1	Height
X_2	Chest diameter
X_3	Shoe size
X_4	Age
X_5	Fasting blood sugar level
X_6	Diastolic blood pressure level

X represents predictors. An n by p matrix, p being the number of predictors and n is the number of observations. The predictors are continuous variables. Y represents responses. An n by r matrix, r is the number of responses. The response is univariate and continuous.

iii) Data Analysis

a) Selection of maximum dimension

Function u.xenv() of the Renvlp library (Lee and Su, 2018) of version 3.5.1 of the R software (R Core Team, 2018) was used to tune optimal number of dimension u at $\alpha = 0.01$for the predictor envelope model (COOK, 2018, Cook et al. 2013 & 2016). This function outputs the dimension chosen by Akaike Information criterion and Bayesian Information Criterion at the default significance level for testing. The results show that the minimum value under AIC is 4921.674 and that of BIC is 4996.940 for number of u=2. This implies that both AIC and BIC agree on 2 sufficient dimension for the fitting of the model. We then fit the model to e**stimate the parameters of the model using**

b) Model Fitting and Parameter Estimation

$$Y = \mu + \eta^T \Omega^{-1} \Gamma^T (X - \mu X) + \epsilon, \sum_X = \Gamma\Omega\Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \text{(Cooketal (2016)}$$

Here, function xenv() of the Renvlp library (Lee and Su, 2018) of version 3.5.1 of the R software (R Core Team, 2018) was used to fit the envelope model in the predictor space (COOK, 2018, Cook et al. 2013 & 2016) using the maximum likelihood. Since u=2 which lies between 1 and 5, the starting value and block wise coordinate descent algorithm in Cook et al was implemented.

The outputs of the function above contain intercept, eta, gamma, omega, sigma x, beta, and asymptotic standard error for elements in beta under the envelope model. Beta equals gamma by inverse of omega by eta. The absolute Z-scores -estimates divided by their asymptotic standard errors for the estimated coefficients estimated by Ordinary Least Squares (OLS) and the Predictor Envelope Method (EM) is computed as a ratio of the asymptotic standard error of the standard regression over the predictor envelopes.($\boldsymbol{R = SE(B_{ols})/SE(\beta_{EM})}$).

The results are as summarized in the table below.

Table 2: Parameter Estimate

Predictor X_i	β_{OLS}	β_{EM}	Z-score for OLS	Z-score for EM	R
Height	51.07	52.03	0.44	0.45	0.99
Chest Diameter	-0.22	-0.12	0.06	0.03	-0.98
Shoe Size	0.40	0.41	0.09	0.09	1.00
Age	-0.11	-0.01	0.07	0.01	1.09
FBS	-0.11	-0.22	0.04	0.08	1.03
DBP	0.12	-0.00	0.11	0.00	1.11

Based on results in table 2, both methods clearly declare variable Height as the only active predictor of weight of pregnant women due to their Z-scores and ratios of their asymptotic standard error.Except for variable DBP, which OLS declares nearly active when EM declares it clearly inactive, both model declared variable Chest diameter, Shoe Size, Age and FBS inactive predictors of weight in pregnant women. It seems then that EM leads to better inference.

The standard error ratios indicate little difference between OLS and EM models for the active variable while the standard errors for the inactive predictors differ materially. We proceed further to use the model to make some estimations or predictions.

c) Prediction

Using test data the function pred.xenv() of the Renvlp library (Lee and Su, 2018) of version 3.5.1 of the R software (R Core Team, 2018) was used to perform prediction in the envelope model in the predictor space (Cook 2018, Cook et al. 2013 & 2016). The results are displayed in the table below.

Table 4: Prediction performances of OLS and Predictor envelope models

	Envelope Prediction		OLS Prediction	
Actual weight (y)	$\hat{Y}$	SE Prediction		SE Prediction
65	66.21	22.29	67.16	22.66
64	68.64	23.04	70.69	23.37
59	56.21	21.52	57.47	21.99
70	66.99	22.60	68.08	22.98
64	59.13	21.69	60.16	22.02

The function above presents a list of estimations inherited from xenv() and the new value of predictors with which to estimate or predict the value of the response variable. This is given as a p-dimensional vector. SE Pred is standard error for a predicted value $\hat{Y}$. Based on the *SE Pred* for both models, it is crystal clear that predictor envelope model does the job better, even with univariate Y. We now look at the measures of error.

6. Discussion of Results

From theAkaike and Bayesian Information Criterion a dimension of two is sufficient for fitting the model. It can be seen from the outputs of the function that the minimum value under AIC is 4921.674 and that of BIC is 4996.940 which fall on number of u=2. This implies that both AIC and BIC agree on 2 sufficient dimension for fitting predictor envelope model. Thus the model will be adequate with two dimensions.

Table 2 shows that both methods clearly declare variable Height as the only active predictor of weight of pregnant women as predictor envelopes gives 52.03 estimate of B, Z-scores of 0.45 which are the greatest values for envelopes in the table. Similarly, the standard estimates are 51.07 for B and 0.44 for z-score which are the greatest value for standard method in the table. The ratio of their asymptotic standard error is 0.99 for Height variable. Except for variable DBP, which OLS declares nearly active with 0.12 beta and z-score 0.11 and ratio of asymptotic standard error of 1.11 when EM declares it clearly inactive with -0.00 beta value and 0.00 z-score, both model declared variable Chest diameter (OLS:B=-0.22,z=0.06; EM: B=-0.12,z=0.03; R=-0.98), Shoe Size (OLS: B=0.40, Z=0.09; EM: B=0.41, Z=0.09: R=1), Age (OLS: B=-0.11, Z=0.07; EM: B=-0.01,Z=0.01: R=1.09) and FBS (OLS: B=-0.11, Z=0.04; EM: B=-0.22, Z=0.08: R=1-03 inactive predictors of weight in pregnant women. The standard error ratios indicate little difference between OLS and EM models for the active variable while the standard errors for the inactive predictors differ materially.

The results in table 3, the predicted values of OLS (67.16, 70.69, 57.47, 68.08, and 60.08) are higher than that of the EM (66.21, 68.64, 56.21, 66.99, and 59.13) model. Likewise, the standard errors of predicted values are higher in OLS (22-66, 23.37, 21.99, 22.98, and 22.02) than those of EM (22.29, 23.04, 21.52, 22.60, and 21.69). These means EM is superior to standard regression method.

7. Conclusion and Recommendations

It can be concluded that height is a predictor of weight. Judging from the prediction estimates one can say that estimates of EM are closer to the true values than those of of OLS. Thus, it can be concluded that parameters are estimated better in EM than in OLS. Predictor envelopes uses smaller dimension than OLS. On the basis of all these predictor envelope method should be used for estimation of parameters for strategic planning and decision taking.

References

[1] Adeyemi, R. (2007) Statistical Monitoring Models and its Application in Medical Administration. *Nigerian Statistical Association.* Scientific Programme and Abstracts/Summary

[2] Akpa, O. M. and Oyejola, B. A. (2006) Statistical Modeling of HIV/AIDS Epidemics: Nigeria's Most Needful Statistical Support for Meeting the MDGS in HIV/AIDS Intervention Initiatives.*Nigerian Statistical Association*, Conference Proceedings

[3] Bing L. (2018). *Sufficient Dimension Reduction*: Methods and Applications With R.CRC Press, 6000 Broken Sound Parkway NW, Suite 300 Boka Raton, FL 33487-2742New York

[4] BondelL, H. D. Beich B. J and Lexin L (2010) Sufficient Dimension Reduction via Bayesian Mixture Modeling *Journal of the Royal Statistical Society, Series B 71:618624*

[5] Cook R.D (2018): *An Introduction to Envelope, Dimension reduction for Efficient Estimation in Multivariate Statistics* (1st Ed), Wiley, U.S.A

[6] Cook, R. D (1998). *Regression Graphics*, John Wiley & Sons, U.S.A

[7] Cook, R. D., Helland, I. S. and Su, Z. (2013). Envelopes and Partial Least Squares Regression. *Journal of the Royal Statistical Society:* Series B 75, 851 - 877.

[8] Cook, R. D., Forzani, L. and Su, Z. (2016) A Note on Fast Envelope Estimation *Journal of Multivariate Analysis* 150, 42-54.

[9] Cox, D. R. (2006) *Principles of Statistical Inference*, Cambridge University Press

[10] Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, Springer Lee M. and Su Z. (2018) Renvlp: Computing Envelope Estimators, R Package Version 2.5. Available at https://CRAN.R-project.org/package=Renvlp

[11] Yin X. and Hilalu H (2015) Sequential Sufficient Dimension Reduction for large p, small n problems. Journal of Royal Statistical Society 77, 4, 879-892