Correlation Scaled Principal Component Regression

K K Singh*[1], Chiranjeevi Sadu[1] and Amit Patel[1]

[1]RGUKT IIIT Nuzvid, Krishna AP 521202, India
krishnasingh@rgukt.in,chiranjeevi@gmail.com, amtptl93@gmail.com

Abstract. Multiple Regression is a form of model for prediction purposes. With large number of predictor variables, the multiple regression becomes complex. It may underfit on higher number of dimension (variables) reduction. Most of the regression techniques are either correlation based or principal components based. The correlation based method becomes ineffective if data contains large amount of multicollinearity, and principal component approach also becomes ineffective if response variables depends on variables with lesser variance. In this paper, we propose a Correlation Scaled Principal Component Regression (CSPCR) method which constructs orthogonal predictor variables having scaled by corresponding correlation with the response variable. That is, the construction of such predictors is done by multiplying the predictors with corresponding correlation with the response variable and then PCR is applied on varying number of principal components. It allows higher reduction in the number of predictors, compare to other standard methods like Principal Component Regression (PCR) and Least Squares Regression (LSR). The computational results show that it gives higher coefficient of determination than PCR, and simple correlation based regression (CBR).

Keywords: Multiple Regression, Principal Components, Correlations, Multicollinearity.

1 Introduction

Regression analysis is very useful in forecasting and prediction and in the field of machine learning as well. It is also used to understand which the predictor variables are related to the dependent (response) variable, and to explore the forms of their relationships [1]. In some circumstances, It is used to find out causal relationship between the independent and dependent variables. Sometimes this can lead to spurious or false relationships [2], for example, correlation does not imply causation. In classical multiple linear regression analysis, problems will occur if the regressors are either multicollinear or if the number of regressors is larger than the number of observations [5].

The earliest form of regression was of least squares method which was published by Legendre in 1805 and by Gauss in 1809. [2]. Gauss published his further work of the theory of least squares in 1821, including a technique of the Gauss– Markov theorem.

The majority of analyses of multidimensional systems are multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) [5][7][8]. In principal component regression (PCR) [6] the first k principal components (PCs) of the predictors X are obtained and used as regressors. The main idea behind PCR is to compute the principal components and then use the first few of these components as predictors in a linear regression model and fitting is done using the classical least squares procedure. If all PCs are used in the regression model, the response variable will be predicted with the same accuracy as with the least square approach. Although PCR can deal with multicollinearity, but it does not directly infer the correlation between the predictor variables and the response variable.

The usual way to construct latent predictors is to obtain the first k principal components (PCs) of the predictors' variables X. This approach is called principal component regression (PCR). It is observed in [4] that in some circumstances where response variable is depending on the predictor variables with lesser variance, PCR gives quite low values of coefficient of determination between the response variable and the predictors.

In this work, a correlation scaled principal component regression (CSPCR) method is proposed. In this method, first we find the correlation of each predictor variable with response variable, then each predictor variable is multiplied by corresponding correlation value, then PCR is applied on varying number of principal components, this way we construct orthogonal predictor variables having scaled by corresponding correlation with the response variable. The scaling of predictor variable neutralizes the effect of predictor having high deviation and low correlation. This method allows higher reduction in the number of predictors, compared to other standard methods like principal component regression (PCR) and least squares regression (LSR).

A correlation based regression (CBR) is also introduced here. It takes the first k predictors in order of non-increasing correlation with response variable. Further the predictors are selected in such a way that the absolute correlation difference between them is greater than a threshold (say 0.02), which reduces the multicollinearity problem up to significant extent.

It is observed that PCR outperforms CBR when the response variable is led by the predictor variables with high variance, but when response variable is highly related with the predictor with low variance, CBR outperform PCR. The proposed method CSPCR outperforms both as it takes neutralized the effect of high variance predictor by scaling it by corresponding correlation with the response variable.

The following sections are organized as follow, section 2 describes background and theory and the next two subsections are notation and algorithm for introduced method. The section 3 explains simulation and computational result.

2 Background and Theory

2.1 Notations

A linear regression line has an equation [3] of the form $\boldsymbol{Y = a + bX}$ where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

$$b = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})}$$

Or equivalently, $b = (X^T X)^{-1}(X^T Y)$

And a is obtained as follows

$a = \bar{y} - b\bar{x}$

Where,

- y_i denotes the observed response for experimental unit *i*
- x_i denotes the predictor value for experimental unit *i*
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit i
- $\bar{x}$ denotes mean of X, $\bar{y}$ denotes mean of Y

The following estimates are considered to evaluate a regression model namely

Goodness of fit/Correlation Coefficient (R^2): $R^2 = SSM/SST$

Where

Regression Sum of Squared Error:

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$$

Total Sum of Squared Error:

$$SST = \sum_{i=1}^{n}(y_i - \bar{Y})^2$$

Multivariate regression is an extension of simple linear **regression**. It is used when we want to predict the value of a variable based on the value of one or more predictor variables.

Multiple regression is an extension of Multivariate and Simple linear regression. Like multivariate regression, higher degree (>1) polynomials are used in multiple regression analysis.

To perform principal components (PC) regression, we find eigen vectors and this way the independent variables transformed to their principal components. Mathematically, it is written $X'X = PDP' = W'X$ where D is a diagonal matrix of the eigenvalues of $X'X$, where X is centered by subtracting its mean. P is the eigenvector matrix of $X'X$, and W is a data matrix (similar in structure to X) made up of the principal components.

Having multicollinearity, two or more of the independent variables are highly correlated, such that one variable can be predicted using another variable. Consequently, rank of $X'X$ becomes lesser than its full column rank structure. Under such situations, some of the eigenvalues of $X'X$ get very close to 0. This problem can be easily sorted out in PCR by excluding the principal components having small eigenvalues.

2.2 Description of algorithm

The correlation based regression (CBR) method is explained as follows:

Step-I: Compute correlation of each predictor with response variable.

Step-II: Sort the predictor variables with respect to non-increasing order of correlation with response variable.

Step-III: Take first k predictor variables obtained from Step-II and apply multiple regression with response variable. In our case k varies from 2 to 10 (that is less equal to one-fourth of total number of predictor (=40), taken in computation in this work).

The correlation scaled principal component regression (CSPCR) method is explained as follows.

Step-I: Compute correlation of each predictor with response variable.

Step-II: Multiply each element in the predictor with the corresponding correlation obtained in step-I.

Step-III: Apply PCR. Take first k principal component as predictors. In our case k varies from 2 to 10 (that is less equal to one-fourth of total number of predictor (=40)).

The computational analysis is explained in section 3.

3 Simulation and Result

For simulating the computation, a synthetic datasets X is used; the data set X is generated with n = 500 samples with m= 40 predictors from specified distributions.

The data set X is generated as follows: the first 10 columns values are taken from a normally distributed numbers with mean 10 and standard deviation=5, the next 10 columns values are taken from an exponentially distributed numbers with mean 5. the next 10 columns values are taken from an exponentially distributed numbers with mean 25.The last 10 columns values are taken from a normally distributed numbers with mean 50 and standard deviation = 25. As 10 columns are taken from same distribution it includes significant amount of multicollinearity. The first 20 column variable contains lesser variance and later 20 columns. Further to randomness $X = X + \Delta$ is computed, where the matrix Δ is random distribution in the range (0, 1).

Furthermore, we generate a response variable as $y = X * a + \delta$ in two different cases;
Case-I: The first 20 elements of the vector 'a' are generated from a uniform distribution in the interval [-1, 1], and the remaining elements of a are 0.

So, y is a linear combination of the first 20 columns of X plus an error term.
Case-II: The last 20 elements of the vector 'a' are generated from a uniform distribution in the interval [-1, 1], and the remaining elements of a are 0.

So, y is a linear combination of the last 20 columns of X plus an error term.
The error term δ is obtained from the distribution N (0, 0.8).

The summary of the centered data set is tabulated in Table-1, where first column value is average summary of 1st to 10th variables, and second column value is average summary of 11th -20th variables and so on.

Table-1: Summary of the data set

Var	01:10	10:20	20:30	30:40
Min.	-2.40882	-1.0954	-0.9205	-2.5838
1st	-0.67356	-0.6565	-0.6908	-0.6525
Median	0.01522	-0.2414	-0.3756	0.0712
Mean	0	0	0	0
3rd	0.76428	0.4477	0.2209	0.6592
Max.	2.02887	4.5123	2.8584	2.3883

The coefficient of Determination (R^2) is used as result of simulation and this coefficient is compared over the different methods; namely CSPCR, PCR, and CBR. The coefficient of Determination (R^2) is averaged for m = 100 iterations. The numerical values of computational results are tabulated in Table.2.

Table 2: Computational Result of Coefficient of determination and Error rate

	#predictors	Case-I		Case-II	
Method	M	R^2	ErrorRate	R^2	ErrorRate
CBR	2	0.416	10.348	0.381	44.543
	4	0.514	9.574	0.573	37.526
	6	0.695	7.707	0.591	37.482
	8	0.840	5.407	0.727	31.571
	10	0.888	4.736	0.744	29.996
PCR	2	0.021	13.782	0.368	44.430
	4	0.043	13.454	0.708	28.515
	6	0.070	13.439	1.000	0.974
	8	0.558	8.785	1.000	0.851
	10	0.857	4.873	1.000	0.765
CBPCR	2	0.454	10.023	0.618	35.342
	4	0.711	6.984	0.802	25.129
	6	0.866	4.783	0.999	1.021
	8	0.941	3.074	1.000	0.719
	10	0.980	1.770	1.000	0.716

The results depicted in Fig-1 and Fig-2 for Case-I and case-II respectively, show that the proposed CSPCR outperforms other standard approaches, and mainly when we select very few number of predictor variables.

Case-I: response variable (y) depends on first twenty variables (having lower spread) of x with random multiplies in (-1, 1)

Case-II: response variable (y) depends on first twenty variables (having higher spread) of x with random multiplies in (-1, 1)

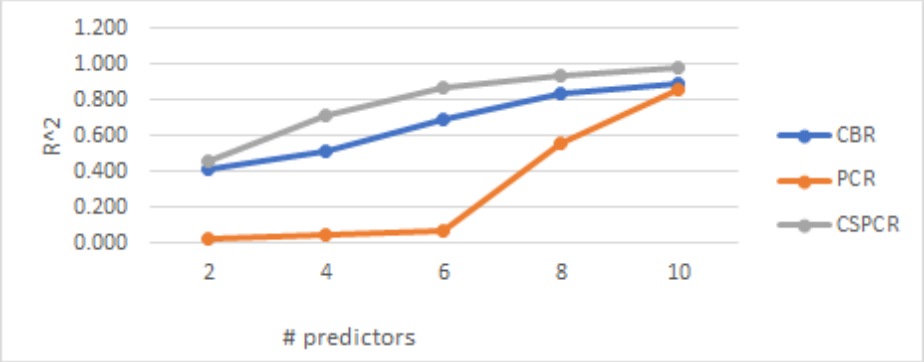

Fig 1: Number of predictors Vs. Coefficient of R^2 in Case-I

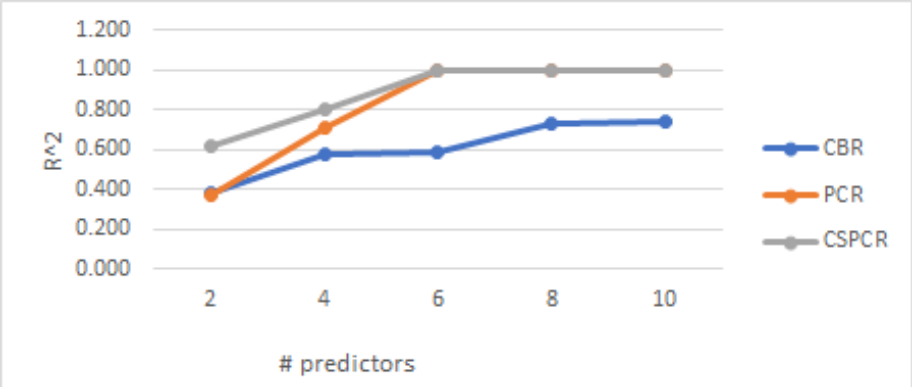

Fig 2: Number of predictors Vs. Coefficient of R^2 in Case-II

1. Conclusion

Problems may occur in multiple linear regression if the regressor variables are highly correlated or number of predictors are less than number of samples. A simple way to get rid of these problems is to apply PCR because it constructs PCs (orthogonal vector) as regressors. But PCs do not take any information of response variable into ac-

count. As it is clear from Fig-1, when y depends on regressors with lesser variations, PCR underperforms than CBR.

The proposed method (CSPCR) constructs the regressors which are scaled by corresponding correlations with the response variable. The simulation study shows that the proposed method allows a significant reduction of the predictor variables compared to PCR and other multiple regression.

References

1. Armstrong, J. Scott (2012). "Illusions in Regression Analysis". International Journal of Forecasting 28 (3): 689. doi: 10.1016/j.ijforecast.2012.02.001
2. Fisher, R.A. (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". Journal of the Royal Statistical Society (Blackwell Publishing) 85 (4): 597–612. doi:10.2307/2341124. JSTOR 2341124
3. M. H. Kutner, C. J. Nachtsheim, and J. Neter (2004), "Applied Linear Regression Models", 4th ed., McGraw-Hill/Irwin, Boston (p. 25)
4. P. Filzmoser and C. Croux, Dimension reduction of the explanatory variables in multiple linear regression, Pliska Stud. Math. Bulgar. 14 (2003), 59–70
5. H. Martens, T. Naes, Multivariate Calibration. Wiley, London, 1993.
6. A. Basilevsky. Statistical factor analysis and related methods: Theory and applications. Wiley & Sons, New York, 1994.
7. Teresa , Araujo, Galvao, , Takashi ,Valeria Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, *CCEN, Caixa Postal 5093, 2001, CEP 58051-970-Joao Pessoa, PB, Brazil.*
8. Kee Siong Ng, A Simple Explanation of Partial Least Squares*, Draft, April 27, 2013*