International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Estimation of Principal Components Regression Coefficients

B. Saikia[1], R. Singh[2]

[1]Research Scholar, Department of Statistics, North Eastern Hill University, Permanent Campus, Mawkynroh-Umshiing, Shillong 793022

[2]Associate Professor, Department of Statistics, North Eastern Hill University, Permanent Campus, Mawkynroh-Umshiing, Shillong 793022

Abstract: ***This paper looks into a method of principal components regression as solution of multicollinearity introducing all the indices of multicollinearity diagnosis. Using this technique, some fairly precise estimates of the coefficients are obtained. This special property of the principal components regression made it superior to the method of ordinary least squares in the presence of multicollinearity in the data. An example is utilized to describe how principal components regression analysis (including all calculating processes) becomes fruitful in case of ill-conditioned explanatory variables.***

Keywords: Least Squares, Multicollinearity, Principal Components Analysis, Tolerance, Variance Inflation Factor and Principal Components Regression

1. Introduction

Estimation of parameter vector by classical linear regression method would have been impossible because of presence of multicollinearity in the data. Popular methods of estimating the parameter vector are: ordinary least squares (OLS), ridge regression (RR), principal components regression (PCR), partial least squares regression (PLSR) and generalized inverse regression (GIR). RR, PCR and GIR are useful while dealing with the presence of multicollinearity in data. But the aim of this paper is to estimate the parameter vector with the help of PCR only for the multicollinear data.

The PCR could be a statistical device that is often suggested as a solution to the multicolliearity problem. Greenberg (1975), Fomby and Hill (1978) and others defined PCR as a method of inspecting the sample data on design matrix for directions of variability and using this information to reduce the dimensionality of the estimation problem. The reduction in dimensionality is achieved by imposing exact linear constraints that are sample specific but have certain maximum variance properties that have make their use attractive. The relationship among p-explanatory variables $X_1, X_2, \ldots, X_p$ is called an exact linear when

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_p X_p = 0 \quad (1)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are constants such that not all of them are simultaneously zero. Let us consider $\lambda_2 \neq 0$ to examine the difference between perfect and nearly perfect multicollinearity. Then (1) is now written as

$$X_2 = -\frac{\lambda_1}{\lambda_2} X_1 - \frac{\lambda_3}{\lambda_2} X_3 - \ldots - \frac{\lambda_p}{\lambda_2} X_p \quad (2)$$

This shows that X_2 is exactly linearly related to other variables. In this situation, the coefficient of correlation between X_2 and the linear combination on the right hand side of (2) is bound to be unity.

The use of principal components (PC) estimators as an estimating procedure in case of multicollinearity is attributed to Kendall (1957) but its most recent proponent has found in McCullum (1970). Kendall's suggestion of artificial orthogonalization can help to alleviate the problem of multicollinearity in regression analysis. Replacement of the correlated explanatory variables by a smaller set of their orthogonal PC can often result in better estimation of the regression parameters than OLS estimation using criterion of mean square error (MSE). Bair et al. (2006) discussed regression problems where the number of predictors greatly exceeds the number of observations and the conventional regression techniques may produce unsatisfactory results. They proposed a technique called supervised principal components (SPC) where the number of variables greatly exceeds the number of samples and it is similar to conventional principal components analysis except that it uses a subset of the predictors selected based on their association with the outcome. Bin et al. (2013) applied SPC to near-infrared and Raman spectral calibration. SPC is similar to traditional principal components analysis except that it selects the most significant part of wavelength from the high-dimensional spectral data, which can reduce the risk of overfitting and the effect of collinearity in modelling according to a semi-supervised strategy. They used three evaluation criteria like coefficient of determination (R^2), external correlation coefficient (Q^2) and root mean square error of prediction to evaluate the performance of each algorithm on both near-infrared and Raman datasets. They considered SPC method might be an alternative method for multivariate spectral analysis.

2. Criteria for Components Selection

A major problem is, 'How does one select component to retain or delete and what are the consequences for each choice'? Usually the number of PCs extracted from the X's is smaller than the number of the X's. Decision for number of PCs to be retained in any particular study is based on some of commonly used criteria like (i) Fomby, Hill and

Johnson criterion (ii) Kaiser's criterion (iii) Cattell's scree-test (iv) Bartlett's criterion (v) Tests of hypotheses criterion and others.

A major decision now has to be made, namely, which component should be retained and which component should be deleted? There are at least two alternatives open for it.

a) Let us delete those components which are unimportant in the sense that they are relatively unsuccessful in explaining the total variability of the group of explanatory variables — obviously deletion of components are associated with small eigenvalues. or
b) Let us delete relatively unimportant components as predictors of criterion variable, i.e. the components having the smallest correlation with criterion variable should be deleted.

Hotelling (1957) has pointed out that, in general, there is no reason why components that are important as far as the predictor variables of a problem are concerned will be highly correlated with the criterion variable in a regression, so criteria (a) and (b) above are likely to lead to different results. If the emphasis is on multicollinearity, as it is here, criterion (a) seems preferable. Jolliffe (1972) has shown with artificial data that deleting those predictor variables associated most strongly with components having low (generally less than 0.7) eigenvalues produces good results in the best subset problem and he selects one variable per component, i.e., that variable having the highest coefficient in the corresponding eigenvector.

3. An Illustration

[Acetylene Data taken from Marquardt and Snee, 1975]: The variables are:

X_1: Reactor temperature in ^{0}c, X_2: Ratio of H_2 to n-heptanes, X_3: Contact time in second, Y: Conversion of n-heptanes to acetylene. The regression of Y on X_1, X_2 and X_3 gives the following results:

Table 1: Coefficients, SE and t-values

Variable	*Coefficient*	*SE*	*t*
X_1	0.127	0.042	3.007
X_2	0.348	0.177	1.967
X_3	-19.022	107.982	-0.176
Constant	-121.270	55.436	-2.188

For N = 16, the value of R^2 is 0.90. R^2 is very high; the coefficient of X_3 is not significant. There is thus a multicollinearity problem. Table 2 displays the mean and standard deviation (SD) of all variables.

Table 2: Mean and SD

	Y	X_1	X_2	X_3
Mean	36.1063	1212.5000	12.4438	0.0403
SD	11.8988	80.6226	5.6620	0.0316

First simple correlations among the explanatory variables are computed and they are $r_{12}^2 = 0.224$, $r_{13}^2 = -0.958$ and $r_{23}^2 = -0.240$. Thus, high correlation between X_1 and X_3 could be the source of trouble. There are several other techniques that are occasionally useful in diagnosing multicollinearity. Tolerance (TOL) and variance inflation factor (VIF) can also be used as a measure of multicollinearity. Larger the VIF, the more troublesome or collinear the explanatory variables is. As a rule of thumb, if the VIF of a variable exceeds 10, that variable will be considered as highly collinear. Again the closer is TOL to zero, the greater the degree of collinearity of that variable with the other explanatory variables. On the other hand, the closer TOL is to 1, the greater the evidence that the explanatory variables are not collinear with the other. Table 3 demonstrates the TOL and VIF and it is observed from this table that TOL $_{X_1}$ and TOL $_{X_2}$ are small (0.082 and 0.081, respectively), VIF $_{X_1}$ and VIF $_{X_2}$ are large (12.225 and 12.325, respectively). These facts also indicate that multicollinearity is present between X_1 and X_3.

Table 3: TOL and VIF

Model	Collinearity Statistics	
	TOL	VIF
1 X_1	0.082	12.225
X_2	0.942	1.062
X_3	0.081	12.325

Table 4: Eigenvalues and PC Coefficients

PC	Eigen values	% of Variance	% of cumulative variance	Coefficient for Principal Component(Correlation coefficient in parenthesis)		
				a_1 (rz1xj)	a_2 (rz2xj)	a_3 (rz3xj)
1	2.060	68.658	68.658	0.675 (0.968)	0.216 (-0.207)	0.706 (0.114)
2	0.899	29.954	98.612	0.294 (0.424)	-0.956 (0.905)	0.011 (0.003)
3	0.042	1.388	100.00	-0.677 (-0.971)	0.011 (0.189)	0.708 (0.145)

The cumulative variance proportion of the first PC Z_1 is 68.658%, the one of the two PCs, Z_1 and Z_2 is 98.612% and one of the three PCs, Z_1, Z_2 and Z_3 is 100.00%. After obtaining the coefficients related to the three PCs to create expressions of three PCs as:

$Z_1 = 0.675X'_1 - 0.216X'_2 + 0.706X'_3$

$Z_2 = 0.294X'_1 - 0.956X'_2 + 0.011X'_3$ and

$Z_3 = -0.677X'_1 + 0.011X'_2 + 0.708X'_3$

Table 5: TOL, VIF, t-values

Model	b_i'	*t*	*P*	*Collinearity Statistics*	
				Tolerance	*VIF*
1 Z_1	0.258	0.997	0.366	1	1
2 Z_1	0.351	0.986	0.342	0.559	1.788
Z_2	.141	0.396	.698	0.559	1.788
3 Z_1	0.163	1.477	0.165	0.547	1.830
	0.88	7.802	0.438	0.558	1.791
Z_2	-0.931	-11.288	0.000	0.971	1.029
Z_3					

We now have obtained the entire standardized partial regression coefficient b_i' of all principal components Z_i in all models (equations) to generate three standardized PCR equations: $\hat{Y}_1' = 0.258C_1$, $\hat{Y}_2' = 0.351C_1 + .141C_2$ and $\hat{Y}_1' = 0.163C_1 + 0.88C_2 - 0.931C_3$. From Table 5, it is observed that the multicollinearity has been reduced. Table 6 shows that their eigenvalues and condition indices (CI) are close to 1. These suggest that all PCs are independent of each other. R^2 is the measure of goodness of fit of linear model and tends to be an overestimate of population parameter. R^2 ranges from 0 to 1. As R^2 is affected by the number of independent variable in the model and sample size, we usually use the adjusted R^2 when comparing the goodness of fit between different linear models. Adjusted R^2 is designed to compensate for the optimistic bias of R^2. Standard error of estimate is the square root of the residual mean squares and measures the spread of the residuals about the fitted line, so it is also a measure of goodness of fit of a linear model. $\hat{Y}_3' = 0.163Z_1 + 0.88Z_2 - 0.931Z_3$ is determined as the best equation, as in Table 7 seen that adjusted R^2 (0.898) and standard error of the estimate (0.3192) of the third standardized PCR equation is the largest and smallest in the three equations respectively. And its F value is equal to 133.159 and it is also highly significant ($P < 0.0005$).

Table 6: Eigenvalues, CI and Variances

Dimension	*Eigenvalues*	*CI*	*% of Variance*		
			Z_1	Z_2	Z_3
1	1.000	1.000	0.50		
2	1.000	1.000	0.50		
1	1.664	1.000	0.17	0.17	
2	1.000	1.290	0.00	0.00	
3	0.336	2.225	0.83	0.83	
1	1.705	1.000	0.15	0.15	0.03
2	1.000	1.306	0.00	0.00	0.00
3	0.965	1.329	0.01	0.03	0.94
4	0.330	2.272	0.84	0.82	0.03

Table 7: Standardized PCR Equations, Adjusted R, SE and F-values

Standardized PCR equation	*Adj* R^2	*SE of estimates*	*F*	*P*
$\hat{Y}_1' = 0.25Z_1$	0.900	0.3165	45.885	0.000
$\hat{Y}_2' = 0.351Z_1 + 0.141Z_2$	0.903	0.3122	70.440	0.000
$\hat{Y}_3' = 0.163Z_1 + 0.88Z_2 - 0.931Z_3$	0.898	0.3192	133.159	0.000

We obtained from Table 4,

$Z_1 = 0.675X'_1 - 0.216X'_2 + 0.706X'_3$

$Z_2 = 0.294X'_1 - 0.956X'_2 + 0.011X'_3$

$Z_3 = -0.677X'_1 + 0.011X'_2 + 0.708X'_3$

which has been applied to the best standardized principal component regression equation: $\hat{Y}' = 0.163Z_1 + 0.88Z_2 - 0.931Z_3$. After having sorted it out we obtained the standardized linear regression equation: $\hat{Y}'' = 0.7662X_1' - 0.4087X_2' - 0.5431X_3'$. The general partial regression coefficient b_i with $b_1' = 0.7662$, $b_2' = -0.4087$ and $b_3' = 0.5431$ and the constant obtained as $b_0 = 25.9684$. Hence, the general linear regression equation is: $\hat{y} = -25.9684 + 0.1131X_1 - 0.8589X_2 - 204.3533X_3$.

4. Discussion and Concluding Remarks

In this paper, a method of PCR is discussed in order to reduce the degree of multicollinearity. And real face of the fact is exposed that b_1= -19.022 is corrected to $b_3^* = 0.5431$ through PCR analysis, which indicates that there is a positive correlation between reactor temperature in ^{0}c and contact time in second. Table 4 shows that the cumulative variance proportion with three PCs goes 100% and namely the best PCR equation $\hat{Y}_3^* = 0.163Z_1 + 0.88Z_2 - 0.931Z_3$ uses all original information. We can perform factor analysis by the standardized partial regression coefficient and also carry out a prediction by means of the general linear regression equation: $\hat{y}$= -25.9684 + 0.1131X_1 - 0.8589X_2 - 204.3533X_3. In multiple linear regression analysis, when there is phenomenon in which results differ from fact, it is usually being suspected about the presence of multicollinearity among explanatory variables. At that time one can use the above method for analyzing. It is thus asserted that the method is computationally effective.

References

[1] Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006), "Prediction by Supervised Principal Components", *JASA*, 101 (473), 119 - 137.

[2] Bin, J., Ai, F., Liu, N., Zhang, Z., Liang, Y., Shu, R. and Yang, K. (2013), "Supervised Principal Components: A New Method for Multivariate Spectral Analysis", *Journal of Chemometrics*, 27 (12), 457 - 465.

[3] Butler, N. A. and M. C. Denham (2000), "The Peculiar Shrinkage Properties of Partial Least Squares Regression", *Journal of Royal Statistical Society (B)*, 62 (3), 585 -593.

[4] Fritts, H.C., Blasing, T. J., Hayden, B.P. and Kutzbach, J.E. (1971), "Multivariate Techniques for Specifying Tree-growth and Climate Relationships and for Reconstructing Anomalies in Paleoclimate," *Journal of Applied Meteorology*, 10 (5), 845-864.

[5] Greenberg, E. (1975), "Minimum Variance Properties of Principal Components Regression", *JASA*, 70 (349), 194 - 197.

[6] Gunst, R.F.; Mason, R.L. (1980), "Regression analysis and its application: A Data- Oriented Approach," New York: Marcel Dekker.

[7] Hotelling, H. (1957), "The Relations of the Newer Multivariate Statistical Methods to Factor Analysis", *British Journal of Statistical Psychology*, 10 (2), 69 -79.

[8] Jolliffe, I. T. (1972), "Discarding Variable in a Principal Component Analysis. I: Artificial Data", *Applied Statistics*, 21 (2), 160 -173.

[9] Kendall, M. G. (1957), "A Course in Multivariate Analysis", Hafner, NY.

[10] Marquart, D.W. and Snee, R. D. (1975), "Ridge Regression in practice", *The American Statistician*, 29 (1), 3 - 20.

[11] McCullum, B. T. (1970), "Artificial Orthogonalization in Regression Analysis", *Review of Economics and Statistics*, 52 (1), 110 -113.

[12] Pasha, G. R., M. A. A. Shah and Ghosia (2004), "Estimation and Analysis of Regression Coefficients When Explanatory Variables are Correlated", *Journal of Research* (*Science*), 15 (1), 33 -39.