Principal Component Regression as a Countermeasure against Collinearity

Chong Ho Yu, Ph.D., Arizona State University, Tempe, AZ

ABSTRACT

There are different approaches to counteract the threat of multicollinearity in regression modeling, such as centered-score regression, orthogonalization, partial least square, and ridge regression. Principal component regression (PCR) is an under-use option because it takes multiple steps to accomplish the goal. This paper will illustrate different steps of performing PCR using the data set compiled by Programme for International Student Assessment (PISA) and a few other databases. Typically, PCR consists of four steps: 1. Principal component analysis (PCA), 2. Principal component regression (PCR) under partial least squares (PLS), 3. Factor analysis, and 4. OLS regression. Initially, PCA is run to verify whether collinear predictors could be combined to form a composite score. Further, the component structure is verified by principal component regression under PLS. While PCA suggests the proper number of principal components by indicating the loadings, PCR makes the same type of suggestion based on the PRESS statistics and variance explained in the model effects. Next, in order to obtain a set of better weights to form a composite index, factor analysis with the varimax rotation is employed. Last, the composite index is used to run the OLS regression model.

INTRODUCTION

The absence of multi-collinearity is essential to a multiple regression model. In regression when several predictors (regressors) are highly correlated, this problem is called multicollinearity or collinearity. When predictors suffer from multicollinearity, using OLS might lead to inflated regression coefficients. These coefficients could fluctuate in sign and magnitude as a result of a small variation in the dependent or independent variables (Fekedulegn, Colbert, Hicks, Schuckers, 2002). Collinearity is problematic when one's purpose is explanation rather than mere prediction (Vaughan & Berry, 2005). Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model. And even if they are not significant, the sum of the coefficient is likely to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal (Leahy, 2001). However, if the goal is explanation, measures other than increasing the sample size are needed. There are many solutions to this problem, such as centered-score regression, orthogonalization, partial least square, ridge regression, and principal component regression (PCR) (Yu, 2008). Space constraints prohibit a thorough discussion of all remedies, and thus in this paper only PCR (Fritts, Blasing, Hayden, & Kutzbach, 1971) is illustrated with archival datasets.

DATA SOURCES

Due to the poor performance of US students in international math and science tests, such as Program for International Student Assessment (PISA), many authors worry that the US lead in science is in jeopardy. President Obama introduced the phrase "Sputnik moment" to characterize this situation (Calmes, 2010; Kornblut & Wilson, 2011). In response to this concern, the author used PCR to determine whether PISA test scores and other economic- and education- related variables are good indicators of economic performance. PISA was administered every three years (2000, 2003, 2006, and 2009) by Organization for Economic Cooperation and Development (OECD) to 15-year old students across various countries. In this example the 2000 PISA test scores (OECD, 2010) were used because it takes a decade to see the return of investment in education.

Common economic- and education- related variables for international comparison, such as real GDP per capita (RGDPL), openness in trade (OPENC), investment share of real GDP (KI), gross secondary school enrollment, the number of scientific and technical journal articles per capita, the number of R&D researchers per capita, and total area of the nation, were included in this study. GDP per capita in 2007 was treated as the dependent variable while all the rest of the above, as well as PISA tests scores, were year 2000 figures and they were used as independent variables.

Population, OPENC, and KI measured in 2000, and real GDP per capita (RGDPL) in 2007, were downloaded from Pen World Table (PWT) maintained by the Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania (2010). Gross secondary school enrollment, the number of scientific and technical journal articles, and the number of R&D researchers collected in 2000, were obtained from the World Bank Group (2010). Although the World Bank's query tool shows that 2008 and 2009 data are available, there are too many missing values in the variables of the focal interest. The latest sufficient data could be found in 2007 only. Also,

University of Pennsylvania has data up to 2007 only. Total area of a nation was obtained from the United Nations (2010); this figure tends to be stable within a decade. In PISA 2000, data for the Netherlands were excluded due to insufficient response rates of students and schools. Last, neither Word Bank nor University of Pennsylvania has sufficient data for Liechtenstein in 2007, and thus this observation was also excluded.

CHECKING ASSUMPTIONS FOR OLS

Regression analysis requires the assumption of linearity, but in GDP per capita, investment share of real GDP per capita, and total area, the extremity of some data values might hinder regression from detecting the hidden pattern. To counteract this problem, a logarithmic transformation was employed to straighten the data. Initially, OLS regression was run in SAS (SAS Institute, 2009) for exploratory purposes. The following is the SAS macros for checking assumptions (Yu, 2010):

```
%macro reg (name, dv, x1, xlast);
ods rtf file="&name..rtf" path="&dir"(URL=none) style=journal;
proc reg data=temp; model &dv =
&x1 &x2 &x3 &x4 &x5 &x6 &x7 &x8/vif;
output out=two
        p=y_hat
        r=y_res;
proc gplot data=two; plot y_res * y_hat;
title "Check homoescedasticity and independence of residuals";
proc univariate normal plot data=two; var y_res;
HISTOGRAM y_res/normal(color=red fill);
PROBPLOT y_res;
QQPLOT y_res;
title "Check normality of residuals";
proc gplot data=temp; plot &dv * (&x1 - &xlast);
title " ";
run;
ods rtf close;
quit;
%mend reg;
```

Although the assumptions for OLS regression, such as normality and independence of residuals, were met, it was found that Luxemburg was an outlier in multiple dimensions, and thus this observation was excluded from regression analysis. In addition, PISA science and math test scores are closely correlated, indicated by their high variation inflation factor (PISA math's *VIF* = 23.98, PISA's science *VIF* = 15.98), and as a result, multicollinearity might have affected the stability of the model. To remediate this problem, PCR was employed as an alternative (SAS Institute, 2009).

PRINCIPAL COMPONENT ANALYSIS

The first step of counteracting multicollinearity is PCA, which is a dimension reduction technique that does not take the correlation between the dependent variable and the independent variables into account. Thus, PCA is considered an unsupervised dimension reduction method. The advantage of this approach is simplicity and therefore this type of dimension reduction is known as "parsimonious summarization" (Maitra & Yan, 2008). The following is the SAS code for running PCA:

```
proc princomp;
var pisa2000math pisa2000sci logki logarea GSSE2000 openc researcherspc
logarticlepc;
```

Table 1 shows the loadings in each eigenvector yielded from PCA of all independent variables. In Table 1, the best loading of each variable is indicated by a bolded number. It is obvious that PISA math and science scores belong to the same principal component while all other variables should be treated as individuals. Hence, it is a reasonable assumption that PISA math and science scores could be reduced to a single principal component. However, in PCA the loading of PISA science score is negative. As a remedy, later factor analysis with varimax rotation was employed in an attempt to obtain a set of better weights.

Table 1. Principal components analysis of all predictors.

Variable	Prin. 1	Prin. 2	Prin. 3	Prin. 4	Prin. 5	Prin. 6	Prin. 7	Prin. 8
PISA math	0.517200	0.152846	-.086731	-.097854	-.154208	0.282416	-.231076	**0.731511**
PISA science	0.496109	0.181937	-.158437	-.127237	-.057514	0.506785	-.043500	**-.646105**
Log(Ki)	0.399195	-.125752	-.363775	0.290402	**0.670608**	-.198618	0.338818	0.064828
Log(Area)	-.224191	0.609222	0.057450	-.077657	0.023531	0.255935	**0.692752**	0.152623
School enrollment	0.151997	0.285523	0.617761	**0.679688**	0.114597	0.041783	-.184547	-.053229
Openness in trade	0.204786	**-.567983**	0.163183	0.263116	-.479925	0.156461	0.532505	0.035068
Researchers/capita	0.195950	-.221648	**0.647426**	-.573322	0.388026	0.022306	0.115124	0.017392
Log(articles/capita)	0.416323	0.321427	0.065762	-.154817	-.358635	**-.729202**	0.138332	-.124807

PRINCIPAL COMPONENT REGRESSION: PRESS

Next, principal component regression (PCR) was employed to verify the suggestion from PCA. In SAS, PCR is not a standalone procedure. Rather, it is an option under partial least square (PLS). Like principal component analysis, the basic idea of PLS is to extract several latent factors and responses from a large number of observed variables. Therefore, the acronym PLS is also taken to mean “projection to latent structure.” In order to invoke PCR, the option “METHOD=PCR” must be specified in the SAS code.

PLS has built-in resampling features, such as leave-one-out (Jackknife) and cross-validation (dividing the data into different subsets for validation). Strictly speaking, leave-one-out and cross-validation are different approaches of resampling, but in PLS both are regarded as cross validation techniques. If the former is used, the SAS code is “CV=ONE”. If the latter is chosen, then the code is “CV=TESTSET”. In this example, leave-one-out was selected because the data set is too small for further partitioning. In addition to the preceding options, the analyst could also choose “CV=RANDOM”, a cross validation technique that randomly resamples subsbets from the original data set. However, if you re-run the procedure again, you will not be able to replicate the same result. Thus, this option is not recommended.

There are different ways of determining the proper number of components to be retained. One of these methods is the predicted residual sum of squares (PRESS). PRESS is the default in SAS, and in the following “CVTEST(stat=press)” is added to the SAS code for demonstration only.

```
proc pls method=pcr cv=one cvtest(stat=press);
model logrgdpl=pisa2000math pisa2000sci logki logarea GSSE2000 openc researcherspc
logarticlepc;
```

Table 2. Press statistics

# of Factors	*Root Mean PRESS*	*Prob > PRESS*	*# of Factors*	*Root Mean PRESS*	*Prob > PRESS*
0	1.052632	0.0320	*5*	0.837301	<.0001
1	0.772102	0.1150	*6*	0.613919	1.0000
2	0.615756	0.4940	*7*	0.648962	0.2810
3	0.69023	0.2800	*8*	0.804963	0.1200
4	0.780427	0.0060			

Minimum root mean PRESS	0.6139
Minimizing number of factors	**6**
Smallest number of factors with p > 0.1	1

In PLS the emphasis is on prediction rather than explaining the underlying relationships between the variables. Thus, PRESS takes mathematical convenience and parsimony into account only. The analyst must make his own judgment to determine whether the result is the best for conceptual explanation. Table 2 shows that six components should be retained according to PRESS. While the model may be more economical than the 7-component model suggested by PCA, it might not be the best conceptual model.

PRINCIPAL COMPONENT REGRESSION: VARIANCE EXPLAINED

In this case the analyst overrides PRESS and considers the alternative: Checking the variance explained in the model effects. In the following SAS code, eight factors are specified in order to show all possible results:

```
proc pls method=pcr nfactor=8;
model logrgdpl=pisa2000math pisa2000sci logki logarea GSSE2000 openc researcherspc
logarticlepc;
```

Table 3 illustrates the percent variation accounted for by different numbers of principal components. If only one principal component is retained (i.e. collapse all eight variables into a single measure), the percent of the variance explained in the model is only 41.23%. Needless to say, this is inadequate. If all eight principal components are treated as independent variables (i.e. no variable reduction is used), the percent of variance explained becomes 100%, but the model might be over-fitted.

Table 3. Percent variation accounted for by principal components.

Number of Extracted Factors	*Model Effects*		*Dependent Variables*	
	Current	*Total*	*Current*	*Total*
1	41.2372	41.2372	53.2414	53.2414
2	24.0250	65.2622	17.1231	70.3645
3	14.6419	79.9041	0.0162	70.3807
4	8.3075	88.2115	0.4853	70.8660
5	5.5401	93.7516	1.4589	72.3249
6	3.4667	97.2184	12.2936	84.6185
7	2.4464	99.6648	2.2881	86.9066
8	0.3352	100.0000	0.0059	86.9125

To locate the optimal point, the number of principal components is plotted against the percent of variance explained, as shown in Figure 1. As expected, going from one principal component to two components results in a substantive gain, but as more and more components results are added, the gain becomes less and less. When the number increases from 7 to 8, the two points form a plateau. Therefore, a seven-variable solution seems to be optimal. Taking both of the results yielded from PCA and PCR into account, it is a logical conjecture that PISA math and science scores could be combined as a single variable, and thus seven variables were included into OLS regression.

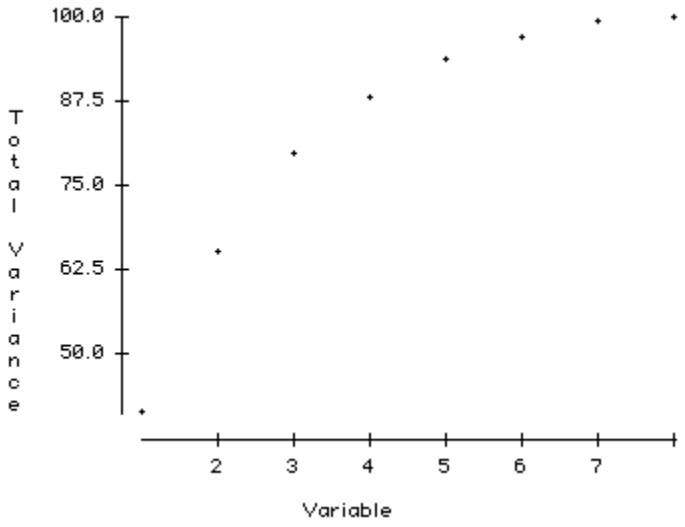

Figure 1. The number of principal components against the percent of variance explained.

FACTOR ANALYSIS WITH VARIMAX

Based on the optimal point of variance explained, the researcher decided to adopt a 7-factor model. As mentioned before, although PCA suggests a 7-component model, some loadings are not desirable. As a remedy, factor analysis with the varimax rotation was employed:

```
proc factor rotate=varimax scree n=7;
var pisa2000math pisa2000sci logki logarea GSSE2000 openc researcherspc
logarticlepc;
```

As indicated in Table 4, the result of factor analysis confirms that of PCA because, based on the factor loadings (highlighted by bolded numbers). PISA math and science scores were considered to belong to the same factor while other variables were suggested to be standalone. Nonetheless, due to the advantage of factor rotation (varimax), factor analysis yields a better set of composite indices for PISA scores because both weights are positive.

Table 4. Factor loadings by varimax rotation.

	Factor1	*Factor2*	*Factor3*	*Factor4*	*Factor5*	*Factor6*	*Factor7*
PISA 2000 math score	**0.92861**	0.08523	0.11160	0.08429	0.20201	0.21593	-0.08811
PISA 2000 science score	**0.95294**	0.03603	0.05449	0.06116	0.24082	0.11456	0.01582
Log(Ki) in 2000	0.39851	0.16218	-0.01141	-0.02081	**0.88981**	0.06581	-0.13502
Log(Area) in 2000	-0.06521	-0.61806	0.16361	-0.17108	-0.25436	0.04028	**0.70098**
Gross sec. school enroll. in 2000	0.11958	-0.02856	**0.98055**	0.09893	-0.00925	0.09427	0.06823
Openness in trade in 2000	0.05378	**0.97504**	0.00393	0.15092	0.09106	-0.02279	-0.12165
R&D researchers per capita in 2000	0.09573	0.17375	0.10191	**0.97078**	-0.01259	0.04677	-0.07423
Log(Sci. articles per capita in 2000)	0.60903	-0.05932	0.17903	0.08221	0.09043	**0.75989**	0.03400

OLS REGRESSION

PISA math and science scores are combined as a composite score using the weights from factor analysis. Table 5 shows the result of the OLS regression model. No violation of assumption, including the absence of collinearity, was found. It shows that when other standard variables for international comparison were taken into account, the number

of scientific and technical journal papers per capita in 2000 appears to be the most significant predictor to GDP per capita in 2007 (p=.0002), and PISA test scores did not substantively contribute to the overall variance explained (R^2 = 0.8691; adjusted R^2=0.7927).

Table 5. Parameter estimates of OLS regression with PISA math and science scores as a single component.

Variable	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$
Intercept	11.81295	1.61832	7.30	<.0001
PISA 2000 math and science score	-0.00030550	0.00090978	-0.34	0.7428
Log(Ki) in 2000	0.40061	0.27889	1.44	0.1764
Log(Area) in 2000	0.03209	0.03933	0.82	0.4304
Gross secondary school enrollment in 2000	-0.00026466	0.00257	-0.10	0.9197
Openness in trade in 2000	0.00048488	0.00143	0.34	0.7403
R&D researchers per capita in 2000	0.00002718	0.00014703	0.18	0.8565
Log(Scientific/technical journal articles per capita in 2000)	0.39280	0.07447	5.27	0.0002

DISCUSSION

In summary, PCR consists of four steps: 1. PCA, 2. PCR under PLS, 3. Factor analysis, and 4. OLS regression. Initially, PCA was run to verify whether PISA science and math test scores could be combined to form a composite score. Further, the component structure was verified by principal component regression under the partial least squares (PLS) procedure. It is important to point out that PCR is not a standalone procedure in SAS; rather, it is offered as an option under PLS. While PCA suggests the proper number of principal components by indicating the loadings, PCR makes the same type of suggestion based on the PRESS statistics and variance explained in the model effects. Next, in order to obtain a set of better weights to form a composite index of PISA math and science scores, factor analysis with the varimax rotation was employed. Last, instead of regressing PISA science and math test scores on the dependent variable directly, the composite index of these two independent variables was used to run the OLS regression model.

When prediction instead of explanation is the research goal, multicollinearity is not a serious threat to the validity of regression modeling. There are many remedies but it is important to emphasize that some of the fixes also aim to making prediction rather than explanation. This example shows that by following PRESS a parsimonious six-component model should be used. However, only PISA math and science test scores were combined as one variable because it is the conviction of the author that all other variables should remain standalone. Nonetheless, PCA, factor analysis, and checking variance explained in PCR substantiate this assertion. As a result, the OLS regression is more meaningful while the problem multicollinearity is solved.

REFERENCES

- Calmes, J. (2010, December 6). Obama calls for new Sputnik moment. *New York Times*. Retrieved from http://thecaucus.blogs.nytimes.com/2010/12/06/obama-calls-for-new-sputnik-moment/
- Center for International Comparisons of Production, Income and Prices (2010). *Pen World Table 6.3*. Retrieved from http://pwt.econ.upenn.edu/php_site/pwt63/pwt63_form.php.
- Fekedulegn, D. B., Colbert, J. J., Hicks, Jr., R. R., & Schuckers, M. E. (2002). Coping with multicollinearity: An example on application of Principal Components Regression in Dendroecology. *Research Paper NE-721, United States Department of Agriculture*. Retrieved from www.fs.fed.us/ne/morgantown/4557/dendrochron/rpne721.pdf
- Fritts, H. C., Blasing, T. J., Hayden, B. P., & Kutzbach, J. E. (1971). Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate. *Journal of Applied Meteorology,10*, 845-864.
- Kornblut, A. E., & Wilson, S. (2011, January 26). State of the Union 2011: 'Win the future,' Obama says.

Washington Post. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2011/01/25/AR2011012504068.html

- Leahy , K. (2001). Multicollinearity: When the solution is the problem. In Olivia Parr Rud (Ed.) *Data Mining Cookbook* (pp. 106 - 108). New York: John Wiley & Sons, Inc.
- Maitra, S., & Yan, J. (2008, November). *Principle component analysis and partial least squares: Two dimension reduction techniques for regression.* Paper presented at Casualty Actuarial Society, Seattle, WA. Retrieved from www.casact.org/pubs/dpp/dpp08/08dpp76.pdf
- OECD. (2010). *Database: PISA 2000*. Retrieved from http://pisa2000.acer.edu.au/
- SAS Institute. (2009). *SAS 9.2*. [Computer software and manual]. Cary, NC: Author.
- United Nations. (2010). *United Nations Statistical Division (UNSD) statistical databases*. Retrieved from http://unstats.un.org/unsd/databases.htm
- Vaughan, T. S., & Berry, K. E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education, 13*(1). Retrieved from www.amstat.org/publications/jse/v13n1/vaughan.html
- World Bank group (2010). *World data bank*. Retrieved from http://databank.worldbank.org/ddp/home.do?Step=1&id=4.
- Yu, C. H. (2008). *Multi-collinearity, variance inflation, and orthogonalization in regression*. Retrieved from http://www.creative-wisdom.com/computer/sas/collinear.html
- Yu, C. H. (2010). *Checking assumptions in regression*. Retrieved from http://www.creative-wisdom.com/computer/sas/regression_assumption.html

ACKNOWLEDGMENTS

Special thanks to Ms. Elizabeth Farley-Metzger for proofreading a portion of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chong Ho Yu, Ph.D.
Arizona State University
PO BOX 612
Tempe AZ 85280
USA
chonghoyu@gmail.com
http://www.creative-wisdom.com/pub/pub.html