An Overview of Remedial Tools for Collinearity in SAS

Chong Ho Yu, Tempe, AZ

Abstract

This paper is an overview of how built-in and customized SAS codes can remediate the problem of collinearity in regression. Collinearity may be caused by (i) too many redundant variables, (ii) the presence of latent variables, (iii) the presence of high-order interaction terms, and (vi) the dependence of variables in a polynomial model. This paper discusses how reducing variables, centering scores, partial orthogonalization, and full orthogonalization should be used in different situations.

Objectives

This paper is an overview of how built-in and customized SAS codes can remediate the problem of collinearity in regression and to clarify some common misconceptions.

The problem of multi-collinearity is often caused by including too many regressors in a regression model. It is a common misconception that stepwise regression enables a researcher to select a subset of variables based upon their relative "importance." Indeed if variables are correlated, the "importance" of the variables are tied to the selection order. Other variable selection criteria such as maximum R-square and Mallow's Cp are recommended instead. Further, if correlated variables indicate latent variables, partial least square procedure is recommended.

Another confusion is the distinction between mathematical dependence and logical dependence. In a regression model involving interaction terms, the interaction variable is highly related to other independent variables. However, the problem of multi-collinearity does not invalidate the regression model. It is because the interaction is only mathematically dependent but not logically dependent on other predictors. A partial orthogonization method or a centered-score regression can be used while an interaction term is present.

A polynomial regression presents a similar confusion. In a polynomial regression the quadratic term, the cubic term, or the quartic term is certainly correlated to the original variable, how can the problem of collinearity be overcome? In this case, a full orthogonized regression model is recommended.

Collinearity

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 multi-collinearity or collinearity. Collinearity means codependence. When variables are related, they are linearly dependent on each other because one can nicely fit a straight regression line to pass through many data points of those variables.

Collinearity is problematic when one's purpose is explanation rather than mere prediction. 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. Thus, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal (Leahy, 2000). However, if the goal is explanation, other measures other than increasing the sample size are needed.

VIF as collinearity diagnostics

Understanding multi-collinearity should go hand in hand with understanding variation inflation. Variation inflation is the consequence of multi-collinearity. In a regression model we expect a high variance explained (R-square). The higher the variance explained is, the better the model is. However, if collinearity exists, probably the variance, standard error, parameter estimates are all inflated. In other words, the high variance is not a result of good independent predictors, but a mis-specified model that carries mutually dependent and thus redundant predictors! Variance inflation factor (VIF) is a common way for detecting multicollinearity. In SAS you can obtain VIF in the following ways:

```
PROC REG; MODEL Y = X1 X2 X3 X4 /VIF
```

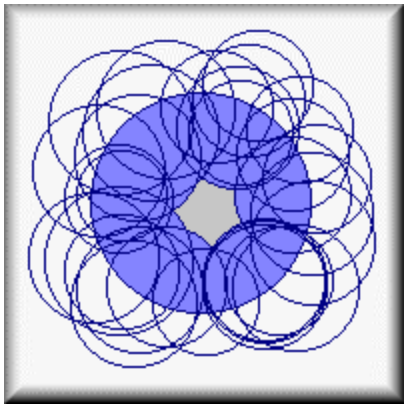
The VIF option in the regression procedure can be interpreted in the following ways:

1. Mathematically speaking: VIF = 1/(1-R-square)
2. Procedurally speaking: The SAS system put each independent variable as the dependent variable e.g.

   X1 = X2 X3 X4
   X2 = X1 X3 X4
   X3 = X1 X2 X4

   Each model will return an R-square and VIF. We can decide which variable to throw out by examining the size of VIF. A general rule is that the VIF should not exceed 10 (Belsley, Kuh, & Welsch, 1980).
3. Graphically speaking: In a Venn Diagram, VIF is shown by many overlapping circles. In Figure 1, the circle at the center represent the outcome variable and all surrounding ones represents the independent variables. The superimposing area denotes the variance explained. When there are too many variables, it is likely that Y is almost entirely covered by many inter-related Xs. The variance explained is very high but this model is over-specified and thus useless. This is a typical problem of too many variables.

Figure 1. Venn Diagram of VIF



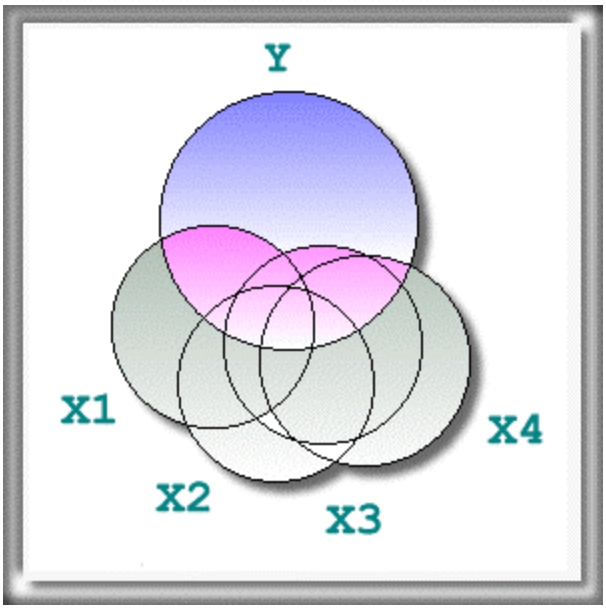
The problem of too many variables

Collinearity happens to many inexperienced researchers. A common mistake is to put too many regressors into the model. When there are too many variables in a regression model, the number of parameters to be estimated is larger than the number of observations. As a result, this model is said to be lack of degree of freedom and becomes over-fitting.

Stepwise regression

One common approach to select a subset of variables from a complex model is stepwise regression. A stepwise regression is a procedure to examine the impact of each variable to the model step by step. The variable that cannot contribute much to the variance explained would be thrown out. There are several versions of stepwise regression such as forward selection, backward elimination, and stepwise.

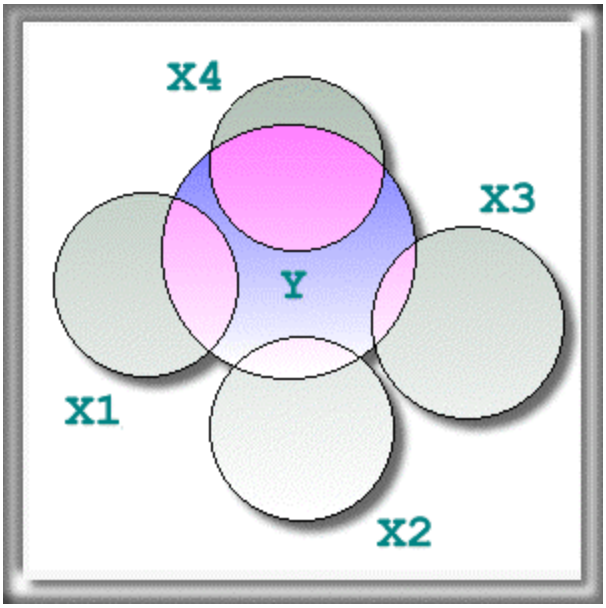
Figure 2. Correlated predictors



However, the above interpretation is valid if and only if all predictors are independent. Collinear regressors or regressors with some degree of correlation would return inaccurate results. Assume that there is a Y outcome variable and four regressors X_1-X_4. In Figure 2 X_1-X_4 are correlated (non-orthogonal). One cannot tell which variable contributes the most of the variance explained individually. If X_1 enters the model first, it seems to contribute the largest amount of variance explained. Then X_2 seems to be less influential because its contribution to the variance explained has been overlapped by the first variable, and X_3 and X_4 are even worse.

Indeed, the more correlated the regressors are, the more their ranked "importance" depends on the selection order (Bring, 1996; Fox, 1991). Nevertheless, we can interpret the result of step regression as an indication of the importance of independent variables if all predictors are orthogonal. In Figure 3 there is a "clean" model, in which the individual contribution to the variance explained by each variable to the model is clearly seen. Thus, it can be asserted that X_1 and X_4 are more influential to the dependent variable than X_2 and X_3.

Figure 3. Uncorrelated predictors



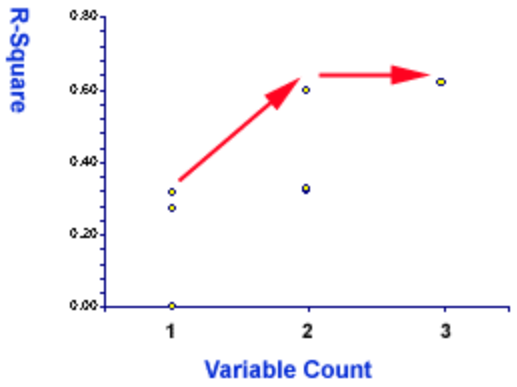
Maximum R-square and Mallow's Cp

Since the purpose of reducing the number of variables is to avoid collinearity, it is absurd to employ a method that is affected by collinearity or some degree of correlation. There are other better ways to perform variable selection such as Maximum R-square (MAXR) and Mallow's Cp. MAXR is a method of variable selection by examining the best of n-models based upon the largest variance explained. Mallow's CP is the total square errors which indicates the lack of fit, as opposed to the best fit by MAXR. Thus, the higher the R-square is, the better the model is. On the other hand, the lower the Cp is, the better the model is. To perform these variable

selection methods in SAS, the syntax is PROC REG; MODEL Y=X1-Xn /SELECTION=MAXR CP

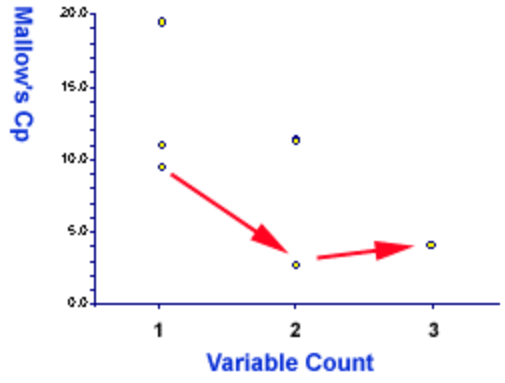
For the clarity of illustration, only three regressors (X_1, X_2, X_3) are used to illustrate MAXR and Cp. The principle illustrated here can be applied to the situation of many regressors. In Figure 4, the x-axis represents the number of variables while the y-axis depicts the value of R-square. In this hypothetical case, it clearly indicates a sharp jump from one variable to two variables. But the curve turns flat from two to three variables (see the arrow).

Figure 4. Plot of number of variables and R-Square



Interestingly enough, in terms of Cp, the full model is worse than the two-variable model. (see the arrow in Figure 5). Nevertheless, although the approaches of MAXR and Mallow's Cp are different, the conclusion is the same: One is too few and three are too many.

Figure 5. Plot of number of variables and Cp



Partial least squares regression

There are other ways to reduce the number of variables such as factor analysis, principal component analysis and partial least squares. The philosophy behind these methods is very different from variable selection methods. In the former group of procedures "redundant" variables are not excluded. Rather they are retained and combined to form latent factors. It is believed that a construct should be an "open concept" that is triangulated by multiple indicators instead of a single measure (Salvucci, Walter, Conley, Fink, & Saba, 1997). In this sense, redundancy enhances reliability and yields a better model.

However, factor analysis and principal component analysis do not have the distinction between dependent and independent variables and thus may not be applicable to research with the purpose of regression analysis. One way to reduce the number of variables in the context of regression is to employ the partial least squares (PLS) procedure. PLS is a method for constructing predictive models when the variables are too many and highly collinear (Tobias, 1999). Besides collinearity, PLS is also robust against other data structural problems such as skew distributions and omission of regressors (Cassel, Westlund, & Hackl, 1999). It is important to note that in PLS the emphasis is on prediction rather than explaining the underlying relationships between the variables.

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. The following is an example of the SAS code for PLS: PROC PLS; MODEL; y1-y5 = x1-x100; Note that unlike an ordinary least squares regression, PLS can accept multiple dependent variables.

The problem of interaction effect

Mathematical dependence and logical dependence

Even if a model is as simple as employing four independent variables, collinearity may still happen when a composite score is included in the model. The following is a typical example:

> GPA = GRE-verbal + GRE-quantitative + GRE-analytical + GRE-total

In the above example, GRE-total is only the sum of all other predictors. Needless to say, GRE-total is strongly associated with those variables. Technically speaking, they are both mathematically and logically dependent. In terms of mathematics, the number of GRE-total is based upon the numbers of all others. In the logical sense, GRE-total is not a new concept.

However, the following model is legitimate though strong relationships exist among predictors:

> GPA = time spent with family + time spent in church + (time spent with family * time spent in church)

The researcher created the last variable because he suspected that GPA is a function of the interplay between family values and Christian work ethics. Nevertheless, in this case they are mathematically dependent but logically

independent. Mathematically speaking, the interaction effect is the product of the first two variables and they certainly have strong numeric relationships. Conceptually speaking, the interaction is considered a new variable and thus it is logically independent from others. In other words, an interaction term does not invalidate a regression model even though the interaction effect is collinear with the two original variables.

Orthogonalization

In spite of its logical independence, we still have to "orthogonalize" the variables to make them mathematically independent. Orthogonality is a state in which the angle between two vectors is 90 degrees. According to Hacking (1992), orthogonality is not only a pure mathematical concept, but also a cultural concept that carries value judgment:

> Normal and orthogonal are synonyms in geometry; normal and ortho- go together as Latin to Greek. Norm/ortho has thereby a great power. On the one hand the words are descriptive. A line may be orthogonal or normal (at right angles to the tangent of a circle, say) or not. That is a description of the line. But the evaluative 'right' lurks in the background of right angles. It is just a fact that an angle is a right angle, but it is also a 'right' angle, a good one. Orthodonists straighten the teeth of children; they make the crooked straight. But they also put the teeth right, make them better. Orthopaedic surgeons straighten bones. Orthopsychiatry is the study of mental disorders chiefly in children. It aims at making the child-normal. The orthodox conform to certain standards, which used to be a good thing (p.163).

Therefore, in the context of regression, orthogonalization can make a "good" regression model. In subject space (vector space), "orthogonalization" can be viewed as a process of subtracting the vector from its projection (Savile & Wood, 1991; Wickens, 1995). In variable space, "orthogonalization" can be explained as a process of finding the residual of the interaction term.

Figure 6 illustrates how a new vector, W, is made by X – Y in vector space. To subtract Y from X, a parallel line of Y is drawn at the end of X. Then a new vector is formed by joining the origin of X, Y and the other end of Y's parallel. In other words, subtraction creates a new vector pointing to a different direction, which is significantly far away from the original vectors. Although X and Y are highly correlated, which is indicated by the small angle between the two vectors, W is uncorrelated to either X or Y. That's why vector subtraction can help to do away with collinearity.

Figure 6. Subtraction in vector space

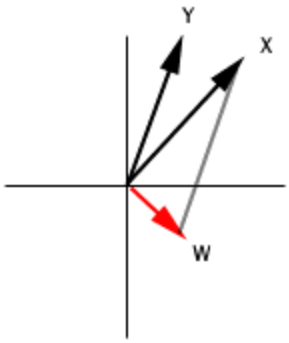
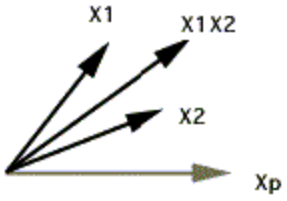


Figure 7 explains orthogonalization by projection. Please keep in mind that this illustration is simplified. In Figure 7, X_1 and X_2 are not strongly related, which is depicted by the wide angel between the two vectors. However, the product of X_1, X_2 is strongly associated with either X_1 or X_2, which is indicated by the proximity between X_1 and X_1X_2, and between X_2 and X_1X_2, respectively.

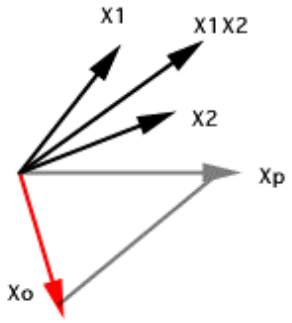
Figure 7. Product vector



To solve this collinearity problem, the first step is to draw a projection of X_1X_2 vector. A projection in subject space is equivalent to the predicted (y-hat) in variable space. In Figure 8, X_1X_2 is the actual vector and Xp is the predicted vector.

After locating the projection, the next step is to create a new vector (new variable), which is orthogonal (not closely related) to X_1 and X_2, but is conceptually equivalent to X_1X_2. By using the subtraction method mentioned above, we can create the new vector Xo. Xo can be viewed as a result of negotiating between what is (X_1X_2) and what ought to be (Xp). Before orthogonalization, there exist a threat of collinearity. After orthogonalization, Xo is far away from X_1 and X_2 and thus collinearity is no longer a threat.

Figure 8. Projection of residual



The SAS code for orthogonalizing the interaction term is as the following. This is a partial orthogonalization method suggested by Burrill (1997):

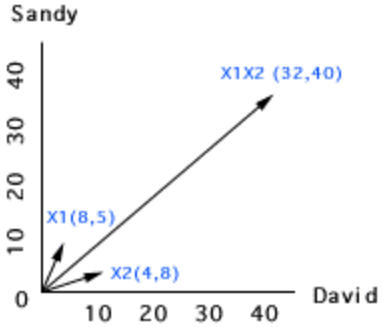
```
X1X2 = X1*X2
/* output the residuals of the interaction term*/
PROC REG DATA=DATA1;
    MODEL X1X2 = X1 X2;
        OUTPUT OUT=DATA2 R=R_X1X2;
/* use the residual as an orthogonalized variable */
PROC REG DATA=DATA2;
    MODEL Y = X1 X2 R_X1X2;
```

Deviation scores

Using centered scores, also known as deviation scores, is another way to avoid collinearity in regression that involves interaction. A centered score is simply the result of subtracting the mean from the raw score (X - X_mean). For the ease of illustration, the following example will use only two subjects and two variables:

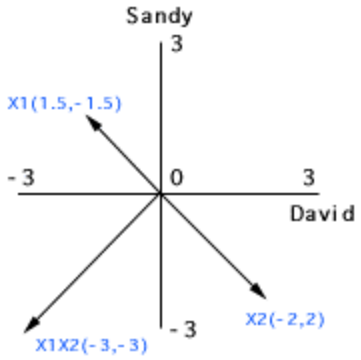
If the raw scores of two subjects are plotted into subject space, there are two short vectors and a very long vector. Two problems are resulted from using the raw scores. First, the scales of X1, X2 and X1 * X2 are very different. Second, there exists collinearity, of course.

Figure 8. Interaction term in vector space



To overcome these problems, one can apply centered scores into the regression model. When one plots the centered scores into subject space, one can find that the scales of all vectors are closer to each other. Further, the interaction term is orthogonal to both centered X1 and centered X2. Hence, collinearity is no longer a threat.

Figure 9. Vectors with centered scores



The SAS code to run a regression with centered scores is as the following:

```
DATA ONE; SET DATA;
        INPUT Y X1 X2;
        PROC MEANS; VAR X1 X2;
        OUTPUT OUT=NEW MEAN=MEAN1-MEAN2;
DATA CENTER; IF _N_ = 1 THEN SET NEW; SET ONE;
        C_X1 = X1 - MEAN1;
        C_X2 = X2 - MEAN2;
        C_X1X2 = C_X1 * C_X2;
        PROC GLM; MODEL Y = C_X1 C_X2 C_X1X2;
```

The preceding code works fine with a small dataset (e.g. a few hundred observations). However, The following revised code is more efficient for a large dataset (e.g. thousands of observations):

```
DATA ONE; SET DATA;
        INPUT Y X1 X2;
        PROC MEANS; VAR X1 X2;
        OUTPUT OUT=NEW MEAN=MEAN1-MEAN2;
DATA CENTER/VIEW=CENTER;
        IF _N_ = 1 THEN SET NEW(KEEP = MEAN1-
MEAN2); SET ONE;
        C_X1 = X1 - MEAN1;
        C_X2 = X2 - MEAN2;
        C_X1X2 = C_X1 * C_X2;
        PROC GLM DATA=CENTER; MODEL Y = C_X1
C_X2 C_X1X2;
```

First, using a view instead of creating a new dataset can save memory space. A view works like a dataset except that it creates the dataset only and only if the data are read. Second, because only the means will be used later, other unused variables can be dropped and only the means are kept. Again, it can save memory space to make the program more efficient.

Polynomial regression

Not all regression models are linear. In some situations the relationship among variables may be non-linear. A classical example is stress-performance relationship. Initially pressure could lead to better efficiency. But if the stress is too intense, performance will decrease due to physical or mental break down.

Another classical example is the relationship between performance and ability. Contrary to popular belief, increasing ability in a discipline or a specific task does not lead to a linear increase in performance. Many teachers are frustrated with the phenomenon that many low achievers do not show improvement in test scores despite tremendous efforts contributed by both teachers and students. It is because low-ability learners do not have the required skills to perform even the basic function. Once they master the basic skills, their performance gain would be proportional to their ability gain. The curve hits an inflection point and turns virtually flat again when the skills are mastered. For example, the score difference in a writing test between a master and a Ph.D. may be minimal. The technical term for this S-shaped curve is ogive (see Figure 10).

Figure 10. Ogive in a polynomial model

In curvilinear cases, polynomial regressions, which involve quadratic, cubic, or quartic terms, should be implemented. Are the quadratic, cubic, quartic and the original variables highly correlated? Yes, it is because the first three are derived from raising power of the original variable. To avoid the problem of multi-collinearity, again you should "orthogonalize" the vectors. In this case a full orthogonalization approach should be used. Gram-Schmidt method is one of the widely used full orthogonalization method but it is difficult to understand and implement.

Another way to orthogonalize the vectors in the regression is to employ PROC ORTHOREG in SAS. This procedure is specifically developed for ill-conditioned data and polynomial model. The orthogonalization method here is Gentleman-Givens transformations. The following example is a labor statistics dataset in the SAS manual. Price level, GNP, unemployment rate, size of armed forces, population, and year are used to predict employment rate. Since the raw variables are strongly correlated and it is believed that the regression model is quadratic, PROC ORTHOREG instead of PROC REG or PROC GLM is used in the estimation.

```
proc orthoreg;  model Employment =
                    Prices  Prices*Prices
                    GNP     GNP*GNP
                    Jobless  Jobless*Jobless
                    Military Military*Military
                    PopSize  PopSize*PopSize
                    Year     Year*Year;
```

Conclusion

There is no single solution to the problem of collinearity. The rationale of using variable selection, latent construct, centered scores, partial orthogonalization, and full orthogonalization must be carefully examined while the researcher encounters various datasets.

Acknowledgement

Special thanks to Dr. Barbara Ohlund and Eldon Norton for reviewing this paper.

References

Belsley, D. A.; Kuh, E.; & Welsch, R. E. (1980). Regression diagnostics : Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

Bring, J. (1996). A geometric approach to compare variables in a regression model. The American Statistician, 50, 57-62.

Burrill, D. (1997). Modeling and interpreting interactions in multiple regression. [On-line]. Available URL: http://www.minitab.com/

Cassel, C.; Westlund, A. H.; & Hackl, P. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. Journal of Applied Statistics, 26, 435-448.

Fox, J. (1991). Regression diagnostics. Newbury Park: Sage Publications.

Hacking, I. (1992). The taming of chance. Cambridge, UK: Cambridge University Press.

Leahy, K. (2000). Personal communication.

Salvucci, S.; Walter, E., Conley, V; Fink, S; & Saba, M. (1997). Measurement error studies at the National Center for Education Statistics. Washington D. C.: U. S. Department of Education.

Saville, D. & Wood, G. R. (1991). Statistical methods: The geometric approach. New York: Springer-Verlag.

Tobias, R. D. (1999). An introduction to partial least squares regression. Cary, NC: SAS Institute.

Wickens, T. (1995). The geometry of multivariate statistics. Hillsdale, NJ: Lawrence Erlbaum Associates, inc.