Vol. 24: 15–18, 2003 | CLIMATE RESEARCH Clim Res | Published June 10

Test for harmful collinearity among predictor variables used in modeling global temperature

David H. Douglass[1,*], B. David Clader[1], John R. Christy[2], Patrick J. Michaels[3], David A. Belsley[4]

[1]Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627, USA
[2]Earth System Science Center, University of Alabama in Huntsville, Alabama 35899, USA
[3]Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia 22904, USA
[4]Department of Economics, Boston College, Chestnut Hill, Massachusetts 02467, USA

ABSTRACT: Lower tropospheric temperature anomalies from the global satellite MSU that have been available since 1979 are unique and play a significant role in the continuing climate debate. A number of investigators have analyzed the MSU data using regression analysis to remove the geophysical effects of volcanoes, El Niño/Southern Oscillation, and solar irradiance in an effort to determine any underlying trend line. In a recent paper Santer et al. (2001; J Geophys Res 106:28033–28059) questioned the validity of such studies, noting that large El Niño events have occurred at the same time as 2 major volcanoes. They calculated a correlation between these 2 variables and claimed that this indicates collinearity, which can adversely affect any regression analyses. We examine the issue of collinearity between the volcano and El Niño/Southern Oscillation signals in the analysis of the MSU data. We do this by using the general tests for collinearity of Belsley. There are 2 tests. The first is for *degrading collinearity* on the data matrix of the predictor variables. If the first test fails, a second test for *harmful collinearity* is performed on the coefficients from any regression analysis. Employing these 2 tests, we find that there is no degrading or harmful collinearity used in the modeling of the MSU temperature anomalies.

KEY WORDS: MSU satellite · Temperature · Collinearity · Regression · Correlation

1. INTRODUCTION

The MSU global satellite temperature anomalies (Christy et al. 2000; updates available from ftp://vortex.atmos.uah.edu/msu/t2tl/) are unique because of their uniform coverage and accuracy, and they play a significant role in the continuing climate debate. These data have been employed in climate assessments at the highest national and international levels (e.g. NRC 2000, Houghton et al. 2001). The influence of geophysical effects such as El Niño/Southern Oscillation (ENSO) and volcanic eruptions on temperature anomalies has been studied by Christy & McNider (1994), Michaels & Knappenberger (2000), and Douglass & Clader (2001) using regression analysis.

Santer et al. (2001) questioned the validity of such studies, noting that large El Niño events have occurred at the same time as 2 major volcanic eruptions (El Chichón [28 March 1982] and Mt. Pinatubo [15 June 1991]), partially canceling their effects. They calculated correlations of the order of 0.4 to 0.5 between these 2 variables and claimed that such 'high' correlations indicate collinearity, which can adversely affect any regression analyses using these data. They state that '[p]revious work in this area has either neglected collinearity effects … or … concluded that they are unimportant.' They cite the work of Christy & McNider (1994), Michaels & Knappenberger (2000) and others as being affected by this problem.

Regression analysis is a very common technique widely used to examine the influence of 1 or more *independent* or predictor variables on a *dependent* variable, for example, the influence of ENSO and volcanic eruptions on air temperature anomalies. This technique,

*Email: douglass@pas.rochester.edu

however, can fail if there is a dependent relation among the predictor variables, which is what Santer et al. (2001) have asserted. An example would be a correlation of 1.0 between 2 of the predictor variables. This example of 1.0 is obvious, but what if the correlation is 0.5 or 0.7? This is characterized as a *near* dependency. Does regression analysis fail? This requires appropriate tests and is the subject of this paper.

There is a difference between correlation and collinearity between variables which must be understood. Correlation and collinearity are discussed in Mosteller & Tukey (1977), cited by Santer et al. (2001). Mosteller & Tukey give many simple examples of collinearity and illustrate the associated problems. However, Mosteller & Tukey do not provide any means for determining the presence of collinearity or, if present, assessing the harm it may cause.

Collinearity is only a part of the general subject of ill-conditioned data. To answer the questions raised, one must study the more general subject. The definitive work on ill-conditioned data is that of Belsley (1991). Belsley begins by showing that many commonly used procedures to detect collinearity, such as correlation, different tests on the correlation matrix, bunch maps, and many others, are not adequate in determining when a regression analysis is harmed. He then introduces diagnostic tests for determining degrading collinearity (DCL) and harmful collinearity (HCL). Collinearity is degrading when there are sufficiently strong near dependencies among the regression variables. If there is no DCL, then any regression analysis may be considered free from collinearity errors. If DCL is present, then a further test, HCL, can be used to determine whether the presence of DCL seriously reduces the reliability of regression estimates based on those data. (The collinearity tests discussed below, although complete, are concise and use terms and definitions from matrix analysis not generally familiar to climate scientists. We suggest that the introductory material in Belsley [1991] be consulted.)

2. BELSLEY COLLINEARITY TESTS

2.1. Test 1: Diagnostic for DCL

DCL is determined by examining the regression variables that are going to be used to 'explain' the observed data. In the example below the regression variables are proxies for volcanic eruptions, ENSO, and solar effects. This test involves only the predictor variables, which are represented as a matrix, **X**.

As Belsley (1991, 1993) shows in great detail, DCL is diagnosed by examining the matrix **X** of predictor variables, including a constant column of 1s if an intercept is present. The diagnostic is based upon the singular-value decomposition of **X**:

$$\mathbf{X} = \mathbf{USV}^{\dagger}$$

where **S** is a diagonal matrix of non-negative numbers called the singular values of **X**, **U** is of the same size as **X** and is column-orthogonal, and **V** is square and unitary. The singular-value decomposition is a standard numerical algorithm found in many libraries of numerical recipes and many computational environments such as Mathematica® or Matlab®. From **S** and **V** Belsley constructs a square matrix π, whose dimension equals the number of regression variables, including the constant, if present. The columns correspond to the regression variables, and the rows correspond to the condition indexes, which are simply the ratios of the singular values to the minimal singular value. Condition indexes in excess of 30 indicate the presence of DCL. The entries in the π-matrix are the variance-decomposition proportions, and their magnitudes help to determine which variables are involved in the collinear dependencies. This π-matrix is now a standard function in statistical packages such as SAS® or TROLL®, and is also available in add-on packages for programs such as Mathematica® (see, for example, Belsley 1993).

Conducting the Belsley test for DCL is quite straightforward:

(1) For each condition index in excess of 30 there exists a collinear relation that is degrading. Inversely, if the largest condition index is less than, say, 10, there is no DCL.

(2) For any condition index larger than 30, 2 variables are involved in its corresponding degrading collinear relation only if their entries in the row corresponding to that index are both greater than 0.5. There are additional details for this part that are not needed here.

2.2. The data matrix X of Douglass & Clader (2002)

We now consider the data matrix **X** used by Douglass & Clader (2002). This is the most recent analysis of the global MSU monthly temperature anomalies using regression analysis. The data used in that analysis is fully described by them and is briefly reviewed and described here. In addition to a constant column of 1s, this matrix contains 4 series:

- AOD (aerosol optical depth) index of Sato (1993; updates available at http://www.giss.nasa.gov/ data/ strataer/) for the volcano. The AOD index is strongly affected by the aerosols given off during a volcanic eruption and is commonly used as a proxy.
- SST 3.4 index (Garrett 2000) for the ENSO. There is a region in the Pacific Ocean designated 3.4. The temperature anomalies (SST 3.4) of this region are

strongly correlated with climate associated with the ENSO effects. This also is frequently used as a proxy.
- Solar irradiance measurements of Lean & Rind (1998). The solar irradiance data of Lean & Rind very clearly show the activity sun spot cycles.
- A linear or trend term.

The first 2 variables are the same as those used in Santer et al. (2001). Michaels & Knappenberger used the first 3, and Douglass & Clader (2002) used all 4.

Table 1 shows the correlation matrix for the data **X** used by Douglass & Clader (2002).

It shows a correlation of 0.415 between SST and AOD, which is consistent with the values reported by Santer et al. (2001).

Douglass & Clader (2002) determined that the influences of volcanoes and El Nino were consistent with the earlier work by Christy & McNider (1994) and by Michaels & Knappenberger (2000). In addition, they were able to determine the solar irradiance constant, $k = 0.11$ K/(W/m^2), which is about twice that expected from a no-feedback Stefan-Boltzmann climate model based upon radiation equilibrium. Also there was a linear trend in the data of 62 mK decade^{-1}. This is also consistent with estimates of Christy & McNider (1994) and Michaels & Knappenberger (2000).

About 93% of the variance was removed in this regression analysis. All results were robust under truncation of the data from either end. The collinearity issues considered in this paper were not considered by Douglass & Clader (2002).

Table 1. Correlation matrix for the data **X** used by Douglass & Clader (2002). SST: SST 3.4 index (Garrett 2000); AOD: aerosol optical index of Sato (1993); Solar: solar irradiance measurements of Lean & Rind (1998); Linear: a linear or trend term

	SST	AOD	Solar	Linear
SST	1			
AOD	0.415	1		
Solar	−0.050	0.025	1	
Linear	−0.062	−0.085	−0.278	1

Table 2. π-matrix of scaled condition indices and variance-decomposition proportions

Scaled condition index	Const.	SST	AOD	Solar	Linear
1	0.029	0.027	0.059	0.001	0.030
1.5	0.008	0.328	0.055	0.253	0.029
1.6	0.011	0.193	0.002	0.618	0.005
2.4	0.013	0.451	0.809	0.045	0.051
4.6	0.939	0.000	0.075	0.083	0.885

2.3. Test 1 on the Douglass & Clader (2002) data X

Table 2 shows the π-matrix (Belsley 1991) of condition indices and variance-decomposition proportions relevant to these data.

This matrix is the bottom-right 5×5 array. As noted by Belsley (1991), when a constant term is included in the model, its presence must be included in any proper conditioning analysis.

The diagnostic test for DCL is now straightforward: (1) The first column contains the condition indices. Since the largest condition index is less than 10 there is no DCL. (2) There is no need for this part of the test since there is no condition index greater than 30. We determine from this test that there is no DCL involving the constant, SST (ENSO), the volcano signals, the solar signal, or a linear trend term. In particular, the 2 important variables, SST and volcano, are not subject to DCL.

3. TEST 2: DIAGNOSTIC TEST FOR HCL

If the data are further used in a regression context, there is another test we can employ to determine whether the DCL is actually strong enough to be considered harmful in the sense of reducing the signal-to-noise ratio of the estimated regression coefficients so as to seriously reduce their reliability and usefulness. For the example of the MSU temperature anomalies considered here, there is no need to do this test, because for collinearity to be deemed harmful, it must first be deemed degrading, and we have just shown above that this is not the case for these data. However, because of the concerns that have been raised by Santer et al. (2001) about the suitability of the conditioning of these data in determining temperature anomalies, we also present the results of this test for this regression context. In addition, the question of collinearity is important beyond this example, and this test should be generally known to researchers in this and related fields.

Table 3 presents the regression coefficients from the Douglass & Clader (2002) regression analysis mentioned above, along with their estimated standard errors, and the signal-to-noise diagnostic relevant to the test for HCL. The signal-to-noise diagnostic in this case is simply the square of the ratio of the coefficient to its standard error. This test differs substantively from the standard test of significance because the resulting statistic is compared to a far more stringent critical value than that of the standard *F*-test.

This Belsley diagnostic test for the adequacy of signal-to-noise is as follows:
- We first determine a critical value defining adequate signal-to-noise. To do this, we pick a test size and a

- level of adequacy for the test. Here we choose a traditional test size of 0.05. The level of adequacy, as explained in Belsley (1991), can be chosen as a value between 0 and 1, where 0 is the weakest and 1 the most (impossibly) stringent level of adequacy. For this purpose we choose a very stringent level of 0.999. Using these parameters, we find from Table 7.9e of Belsley (1991) that the critical value for adequate signal-to-noise for individual parameters with 255 degrees of freedom is 24.8. (If one reduces the degrees of freedom to 255/4 the value is 26.1.)
- We next calculate the diagnostic test values for the estimated signal-to-noise for each regression coefficient, which for individual coefficients is the square of the ratio of the estimated coefficient to its estimated standard error. These are given in row 3 of Table 3. Baring the completely insignificant intercept estimate, we see the smallest of these estimated signal-to-noise figures is 27.
- Comparing these figures to the critical value of 24.8 (or 26.1), we see that all coefficients except the intercept possess adequate signal-to-noise, this being particularly true for the 2 variables of greatest interest, SST (ENSO) and volcano, both of which have values in excess of 300.

Table 3. Regression coefficients (Douglass & Clader 2002)

	Const.	SST	AOD	Solar	Linear
Regression coefficient	−0.013	0.145	−3.8	0.101	0.00643
Standard error	0.017	0.008	0.2	0.018	0.00124
Signal-to-noise	0.6	327.7	361.0	31.4	27.0
Degrees of freedom	256				

We conclude that there is no evidence of HCL affecting the results of using the ENSO and volcano data in a regression determining MSU temperature anomalies. Indeed, to the contrary, we find these data to be very well conditioned and highly suitable to determine the regression coefficients.

4. CONCLUSION

We have examined the issue of collinearity or ill-conditioning in the volcano and El Niño/Southern Oscillation signals in a study of the MSU global satellite temperature anomalies using the 2 rigorous collinearity tests of Belsley (1991). The first test on the predictor variables showed no degrading collinearity. Even though the first test is sufficient, we performed the second test for harmful collinearity on the constants from the regression analysis. We found no harmful collinearity as has been suggested by Santer et al. (2001)

We have answered the question of collinearity of the chosen regression variables. However, this does not answer other questions such as whether or not these variables are the most relevant choice for climate-forcing parameters or if there are other processes that should be considered. These questions are appropriate for later discussions.

Acknowledgements. This work was supported in part by the Rochester Area Community Foundation. Note. The data used in this paper are available upon request from D.H.D.

LITERATURE CITED

Belsley D (1991) Conditional diagnostics: collinearity and weak data in regression. Wiley Series in Probability. John Wiley, New York

Belsley D (1993) Econometrics: a package for doing econometrics in Mathematica. In: Varian HR (ed) Economic and financial modeling with Mathematica. Springer-Verlag, New York

Christy JR, McNider RT (1994) Satellite greenhouse signal. Nature 367:325–367

Christy JR, Spencer W, Braswell WD (2000) MSU tropospheric temperatures: dataset construction and radiosonde comparisons. J Atmos Ocean Tech 17:1153–1170

Douglass DH, Clader BD (2002) Determination of the climate sensitivity of the earth to solar irradiance. Geophys Res Lett 29 (1029/2002GL015345) 33-1–33-4

Garrett D (2000) Monthly index values (SST). Climate Prediction Center, Boulder, CO. Available from http://www.cpc.ncep.noaa.gov/data/indices/index/html

Houghton JT, Ding Y, Griggs DJ, Noguer M, Linden PJ, Dai X, Maskell K, Johnson CA (eds) (2001) Climate change: 2001—the scientific basis. Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

Lean J, Rind D (1998) Climate forcing by changing solar radiation. J Clim 11:3069–3094

Michaels PJ, Knappenberger PC (2000) Natural signals in the MSU lower tropospheric temperature record. Geophys Res Lett 2:2905–2908

Mosteller F, Tukey JW (1977) Data analysis and regression. Addison-Wesley, Boston

NRC (2000) Reconciling observations of global temperature change. National Research Council, Panel on Reconciling Temperature Observations. National Academy Press, Washington, DC

Santer BD, Wigley TML, Doutriaux C, Boyle JS and 6 others (2001) Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. J Geophys Res 106:28033–28059

Sato M, Hansen J, McCormick MP, Pollack JB (1993) Stratospheric aerosol optical depths, 1850–1990. J Geophys Res 98:22987–22994

Editorial responsibility: Chris de Freitas, Auckland, New Zealand

Submitted: December 3, 2002; Accepted: March 26, 2003
Proofs received from author(s): May 8, 2003