Multicollinearity between the solar proxy and a linear trend in an ordinary least squares regression – Applicable to Mesospheric temperature measurements

T. Wynn, V. B. Wickwar

Abstract

When doing regression multicollinearity between model variables can be a problem. This is a problem for time and solar coefficients for data sets of mesospheric temperatures spanning one solar cycle or less. This paper focuses on the problem of multicollinearity between the linear term and the solar term in an ordinary least squares regression (OLSR). The multicollinearity between those two terms will change according to the phase of the solar cycle. If solar maximum occurs in the middle of the second half of the data set there is significant negative correlation between them. Conversely, if solar maximum occurs in the middle of the first half of the data set there is significant positive correlation. The optimal phase of the solar cycle relative to the data is for solar max or solar min to occur in the time center of the data set. In that particular case the correlation between the linear and solar coefficients is minimized. When the data set spans approximately 1.3 solar cycles or greater then multicollinearity between the time coefficient and solar coefficient is not an issue. The degree of multicollinearity is independent of the magnitude of the solar response and cooling rate.

1. Introduction

There is compelling evidence that the earth's climate is undergoing long-term changes, and there is a strong consensus among scientists that this is largely due to anthropogenic influence. It has been shown that increases in the level of carbon dioxide cause the lower atmosphere (troposphere) and middle atmosphere (stratosphere and mesosphere) to react differently: the lower atmosphere warms and the middle atmosphere cools. Further, the temperature change in the middle atmosphere is expected to be about ten times greater than that in the lower atmosphere (Fomichev, et al., 2007). Hence, many scientists are looking for evidence of anthropogenic influence on atmospheric temperatures in the middle atmosphere. Information about how atmospheric temperatures are evolving on decadal time scales, as well as seasonally, and to external influences such as solar variability is often extracted using ordinary least squares regression (OLSR). If each measurement is unbiased and uncorrelated then this technique provides the best linear unbiased estimator (BLUE). Looking at it a different way, a column of data, temperatures in this case, is projected onto a column space of independent variables as to minimize the variance of the residuals. If the relevant independent variables are included in the model then OLSR minimizes what the model cannot account for. However, if there is a high degree of correlation between explanatory variables interpreting the results is less than straightforward, though the coefficients are still BLUE. How multicollinearity between explanatory variables affects the results of an OLSR needs to be understood and considered in the final interpretation of the results.

2. The problem

OLSR on atmospheric temperatures generally includes the following explanatory variables: annual oscillation and semiannual oscillation, linear trend, and a solar proxy representing changes in solar input. It might also include information about the quasi-biennial oscillation, or short-term effects such as changes in atmospheric optical depth due to volcanic eruption. Consider the following model,

$$T_i = w + b\cdot t_i + s\cdot F107_i + A_1\sin(2\pi\cdot t_i) + A_2\cos(2\pi\cdot t_i) + B_1\sin(4\pi\cdot t_i) + B_2\cos(4\pi\cdot t_i) + \varepsilon_i, \quad (1)$$

where T_i is the temperature at time t_i , b is the linear trend coefficient, s is the solar response coefficient, w is the intercept; from the coefficients A_1 and A_2 the amplitude and phase of the annual oscillation can be extracted; the same is true for the semiannual oscillation coefficients B_1 and B_2 ; ε_i is the residual and F107 is the solar proxy data, in this

case the 10.7 cm radio flux in solar flux units (1 sfu = 10^{-22} W m^{-2} Hz^{-1}), which is sometimes used as a proxy for changes in UV intensity. The following explanatory variables form a column space onto which T is

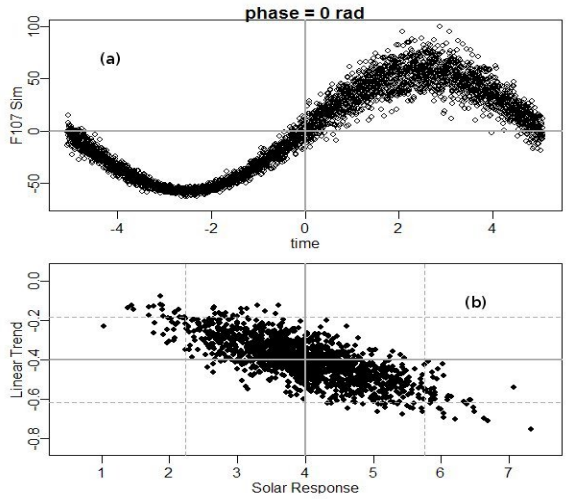

Figure 1: (a) The simulated F10.7 proxy and (b) comparison of the solar response and time coefficients. The solar response coefficients was multiplied by 2*57.6 sfu to puts the solar coefficient on a scale of K.

projected: (1) t, (2) f107, (3) sin(2π·t), (4) cos(2π·t), (5) sin(4π·t), (6) cos(4π·t), and a column of 1's for the intercept. Also, for simplicity, time is adjusted so that t=0 occurs exactly in the time center of the data set. Under ideal conditions the independent variables form an orthogonal column space, in which case there would be no need to consider multicollinearity. Obviously (3) is orthogonal to (4), and (5) is orthogonal to (6), and multicollinearity between any of the sine and cosine terms with the other periodic terms is minimal. So, for examining multicollinearity a simplified model may be considered,

$$T_i = 0 + b \cdot t_i + s \cdot F107_i + \varepsilon_i \,, \qquad (2)$$

where T_i, t_i, b, ε_i are as indicated above. The solar proxy (F107), time (t), and temperatures (T) have zero mean, which allows the regression to be forced through zero, indicated by the 0 on the right hand side of (2). For this analysis, time is in years, making one day equal to (1/365) years, or ~ 2.74×10^{-3} years. To further simplify, a sine function with amplitude of 57.6 sfu and angular frequency of 0.0986 rad/year (a period of ~ 10.14 years) was used in place of the 10.7-cm solar proxy; the phase of the solar function is referenced to the time center (t = 0) of the data set. (See Figure 1a.)

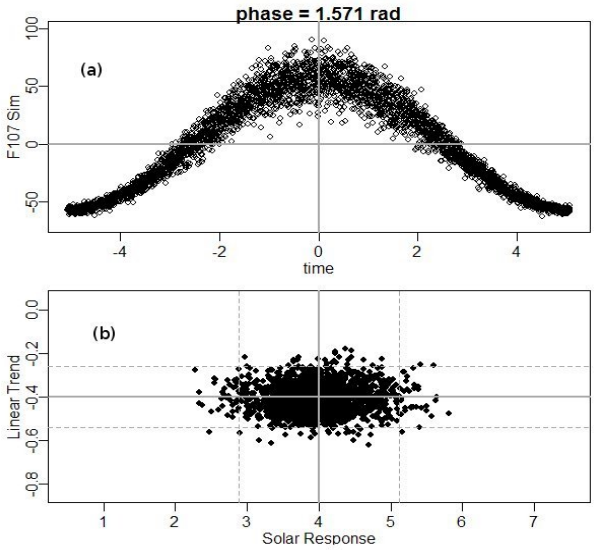

Figure 2: Same as Figure 1 except for a solar phase of 90°.

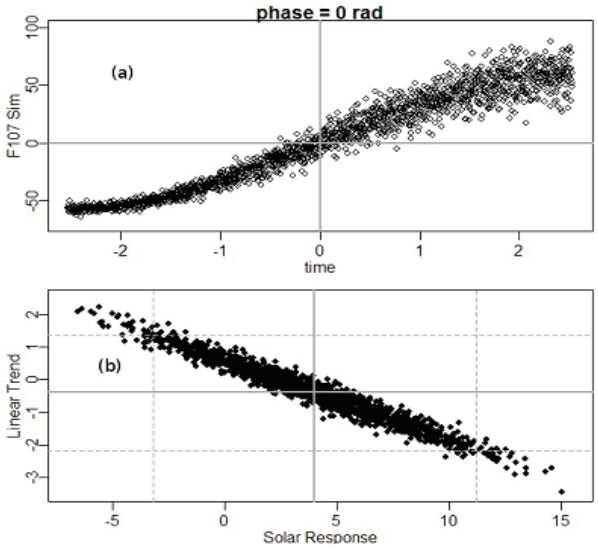

Figure 3: Same as Figure 1 but for one half solar cycle.

3. Strong multicollinearity: one solar cycle

There is a simple way to test for the presence of multicollinearity. After doing an initial regression the residuals and predicted values are obtained. Then, supposing the data set has n data points, n of the residuals are selected with replacement,

meaning that any given residual may be selected more than once or not at all. The selected residuals are added to the predicted values and the regression is repeated. From this new regression slightly different regression coefficients are obtained. This process is repeated about 1000 times and from the set of coefficients obtained a distribution may be inferred for each estimator. This process is known as bootstrapping and has the advantage of avoiding assumptions about the underlying distribution. By plotting one set of coefficients against another the effect of multicollinearity becomes apparent. The presence of significant multicollinearity will create a pattern similar to that in Figure 1b, which shows coefficients from 1500 bootstrapped regressions done on a time series of temperatures having a cooling rate of -0.4 K/year and a 4 K/(max – min) solar response between maximum and minimum. The pattern is that of a bivariate normal distribution.

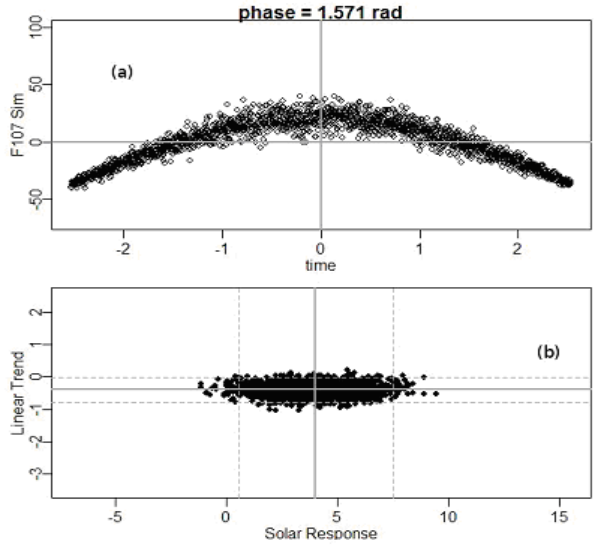

Figure 4: Same as Figure 2 except for a data set spanning one half of a solar cycle. (The mean as been subtracted from the solar signal.)

In the case where there is no multicollinearity the confidence intervals indicate, to a specified level of confidence, the region that presumably includes the true value corresponding to the estimator. Ideally the range of values of one coefficient would say nothing about the others. But if the pattern of coefficients is bivariate then interpretation is more involved. The pattern in Figure 1b indicates possible values that could be obtained from any given regression. The gray dashed lines indicate the 2σ or 95% confidence intervals for the time and solar coefficients. They indicate with 95% confidence (based on the regression) that the true value of the cooling rate is between -0.2 and -0.6 K/year and the true value of the solar response is between 2.2 and 5.7 K from solar max to solar min. But because of the correlation between them, joint inferences cannot be freely made. For example, it would be highly unlikely that the true value of the solar response is 2.5 K and the true value of the time coefficient is -0.6 K/year. Those two values when taken together are outside the elliptical region covered by the time and solar coefficients jointly. It sounds rather counter intuitive, but if our interpretation is constrained by the results then that is how it must be stated. But if it is inferred that the true value of the solar response is 4.7 K then according to that conditional, the cooling rate is between -0.54 to -0.40 K/year with 95% confidence, which is much narrower than the overall spread. If it is inferred that the cooling rate is -0.3 K/year then the solar response is between 2.7 and 3.9 K to the same level of confidence. One cannot make specific inferences about one coefficient without making inferences about the other.

Each data point can also be thought of as a possible mean value for the solar and time coefficient from a given site. Assuming that the temperature data at every site has a -0.4 K/year cooling rate and a 4 K solar response then the distribution shown in Figures 1 and 2 are possible OLSR time and solar coefficients from 1500 data collection sites. Since the standard errors (SE's) will be essentially the same for each point (all other things being equal) they each have a bivariate distribution similar to the overall pattern but centered at their own mean value. Using Figure 1 as an example, if the time coefficient at a given site is higher than the actual cooling rate then the solar coefficient is likely to be too low. But if the time coefficient is lower than the actual cooling rate then the solar coefficient is likely to be too high.

If interpretation is restricted to statements about the range of possible values then multicollinearity is not problematic since there would be a significant amount of overlap in the confidence intervals from each of the sites. But we usually don't know what the actual mean values are, and therefore it's difficult to know if the results from any given site

has a high/high or low/high tendency.

4. No multicollinearity

The case of no multicollinearity occurs when the solar phase angle is π/2 or 3π/2 radians. This is when solar max or min occurs in the time center of the data set. Figures 2 and 4 show cases with no multicollinearity for data sets spanning 1 and 0.5 solar cycles, respectively, and a phase angle of π/2. There is no apparent correlation between outcomes. The true value of the solar (or time) coefficient might be high or low, but this says nothing about the value or confidence interval of the other coefficient. Also, the overall spread is narrower than when multicollinearity is present. The standard deviation of the time and solar coefficients each increased 59% from Figure 4 to Figure 3. But going from Figure 3 to Figure 4 the standard deviation of the solar coefficients increased 96% and that of the time coefficients 327% respectively.

In the case of extreme multicollinearity, shown in Figure 3, the possible values of the true solar and time coefficients are unacceptably imprecise. In cases like this, alternative methods of regression should be considered. However, leaving out the solar proxy variable should not be considered, as it can introduce significant bias in the time coefficient if there is a true solar temperature response in the temperature data.

5. The source of Multicollinearity

The reasons for this response can be seen more clearly in the equation for the standard error of a regression coefficient

$$SE_{bk} = \frac{s_e}{\sqrt{(1-R_k{}^2)TSS_k}}, \qquad (3)$$

where s_e is the standard error of the residuals, R_k^2 is the coefficient of determination from regressing the k^{th} variable on the other variables, and $TSS_k = \Sigma(X_{ki} - \underline{X}_k)^2$, where $\underline{X}$ is the average. The factor $(1 - R_k^2)^{-1}$ is called the variance inflation factor (VIF). Because there are only two explanatory variables in (2), each with zero mean, the coefficient of determination becomes the square of the correlation between them, and $TSS_k = \Sigma(X_{ki})^2 = |X_2|^2$. Rewriting equation (3) for the solar and time coefficient we get

$$SE_t = \frac{s_e}{\sqrt{(1-\rho_{s,t}{}^2)|t|^2}} \qquad (4)$$

$$SE_s = \frac{2A\,s_e}{\sqrt{(1-\rho_{s,t}{}^2)|s|^2}} \quad . \qquad (5)$$

where ρ_{st} is the coefficient of correlation between the solar and time variables and $|s|^2$ and $|t|^2$ are the square of the magnitudes of the solar and time independent variables respectively. Equation (5) was multiplied by (2A) to put it on a scale of $K/(solar_{max} - solar_{min})$, where A is the amplitude of the solar proxy; doing this makes the SE of the solar coefficient independent of the amplitude of the solar proxy and therefore applicable to any solar proxy one would elect to use. With the standard errors written in this form it is easier to see how the interaction of the two independent variables and the length of the data set influence the standard error. A high degree of correlation between the solar proxy data column and time column in (2) creates a high standard error but a longer data set has a larger $|s|^2$ and $|t|^2$, which lowers it.

6. Conclusions

In most cases not much can be done about multicollinearity. But if present it should be understood. The confidence interval for the time coefficient in Figure 1 still spans -0.6 to -0.2 K/year, and the confidence interval for the solar coefficient spans 2.2 to 5.8 K. However, joint inferences are constrained to the elliptical region covered jointly by the bootstrapped coefficients. The tendency between them is, for a solar phase angle of 0 radians, a (low cooling trend)/(high solar response) and (high cooling trend)/(low solar response). For a solar phase angle of π radians the tendency is reversed, high/low and low/high. This can become relevant when comparing results between data sets. Because the true values of the coefficients are unknown, the tendency is also unknown. But if there is no multicollinearity then the value of one coefficient says nothing about the value of the other.

Even with strong multicollinearity between the

explanatory variables the coefficients from OLSR are still BLUE. The consequences are in the interpretation of the data. If a line is drawn through the length of the data (the first principle component along the elliptical spread) in Figure 1b the slope of the line does not depend on the amplitude of the coefficients or their SE's. It depends only on the phase of the solar cycle.

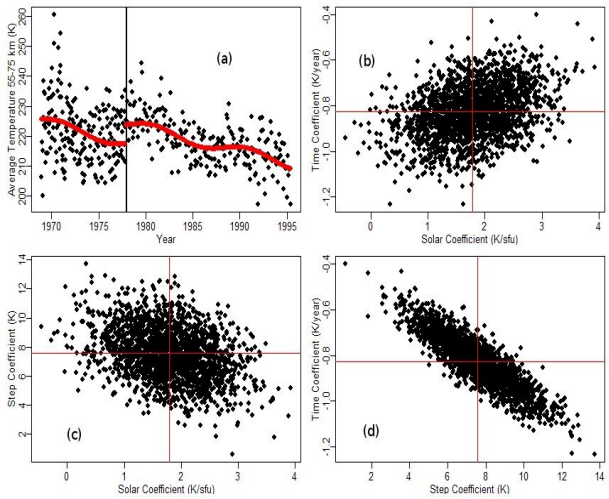

Figure 1: (a) Shows the temperatures from the Volgograd site from January 1969 to September 1995. The vertical line shows when a major sensor change occurred. The solid vertical line in (a) shows the predicted values from an OLSR that included a linear trend, solar proxy, and step function. (b) The bootstrapped coefficients for the solar proxy and time coefficient. Very little multicollinearity is present. (c) The same for (b) but for the coefficients for the step function and time coefficient. Also, very little multicollinearity is apparent. (d) The same as (b) and (c), but for the step function coefficients and linear trend coefficients considerable multicollinearity is present.

The results found here for the interaction of the time and solar data coefficient can be generalized. If there is a large temperature perturbation near the beginning or end of the data set, then multicollinearity is very likely and should be taken into consideration. Any temperature perturbation that goes through one cycle or less over the span of the data set should be considered for possible multicollinearity, e.g, step functions used in regressions on rocketsonde temperatures. Temperatures from several sites (Ryori Japan [Keckhut and Kodera et al., 1999], US rocketsondes in North and South America [Keckhut *et al.*, 1999], and Volgograd Russia [Kubicki *et al.*, 2006]) span several decades. Over that time instrumentation changes occurred that might have introduced bias into the temperatures. This is sometimes accounted for by adding a step function to the OLSR. But this creates an intractable multicollinearity problem between the time coefficient and the step function coefficient. The phase of the step function (taken to be where it goes from its low to its high value, assuming there is only one step) cannot null out. The result is a high degree of correlation between the bootstrapped linear trend coefficients and the step function coefficients, resulting in a pattern very similar to that in Figure 1b. For example, temperatures from the Volgograd site (Figure 5a [Kubicki *et al.*, 2006]). have an instrumentation change a third of the way through their data set. Figure 5(b, c, and d) are multicollinearity plots between the solar, time, and step function coefficients. There is strong multicollinearity between the step function and time coefficients. However, if the magnitude of the step function is not important then joint inference need not be made. If it is, then multicollinearity needs to be considered.

When multicollinearity is present in the data the following difficulties arise. (1) The SE's of highly correlated variables will be much greater than when uncorrelated. (2) Inferences about the actual value of one coefficient must be made jointly with coefficients it is correlated with. (3) Because the actual values of the linear estimators are unknown comparing results from different sites is problematic because the high-high, low-high, or low-low tendency cannot be easily discovered. (3) Multicollinearity between the time coefficient and step function is only problematic if the magnitude of the step function is important.

This analysis of multicollinearity was prompted by the analysis of 11 years of Rayleigh-lidar mesospheric temperatures from USU. A simple OLSR analysis of the data from the upper mesosphere produced a time coefficient of –1 K/year and no dependence on solar input. These results did not seem right—the magnitude of the time coefficient was much bigger than predicted and inferred from the other data at slightly higher altitudes, while the solar dependence was much smaller than inferred from data from slightly higher altitudes. This simulation shows that the results could have arisen from multicollinearity. The best

way to avoid an erroneous result from multicollinearity is to extend the data set over more years.

Acknowledgments: This research has been partially supported by Utah State University and the Rocky Mountain Space Consortium.

References

Fomichev, V. I., Jonsson, A. I., de Grandpré, J., Beagley, S. R., McLandress, C., Semeniuk, K., Shepherd, T. G., 2007, "Response of the Middle Atmosphere to CO2 Doubling: Results from the Canadian Middle Atmosphere Model," *Journal of Climate*, vol. 20, Issue 7, p. 1121.

Keckhut, P., Schmidlin, F. J., Hauchecorne, A., Chanin, M.L., 1999, "Stratospheric and mesospheric cooling trend estimates from u.s. rocketsondes at low latitude stations (8°S-34°N), taking into account instrumental changes and natural variability," *Journal of Atmospheric and Solar-Terrestrial Physics*, Volume 61, Issue 6, p. 447-459.

Keckhut, P., Kodera, K., 1999, "Long-term changes of the upper stratosphere as seen by Japanese rocketsondes at Ryori (39 deg N, 141 deg E)," *Annales Geophysicae*, vol. 17, Issue 9, pp.1210-1217.

Kubicki, A., Keckhut, P., Chanin, M.L., Hauchecorne, A., Lysenko, E., Golitsyn, G.S., 2006, "Temperature trends in the middle atmosphere as seen by historical Russian rocket launches: Part 1, Volgograd (48.68°N, 44.35°E)," *Journal of Atmospheric and Solar-Terrestrial Physics*, Volume 68, Issue 10, p. 1075-1086.