Lab 7
Due Nov 3

Multiple regression using a year-to-year predictor (ENSO) and a gradual trend predictor.

Use two predictors to describe to what degree the year-to-year climate forcers (e.g. ENSO) affect the climate at a given location and season, compared with gradually changing climate forcers (e.g. carbon dioxide concentration, or just the year itself used to indicate time).

How to do a multiple regression in Excel:

We are going to do this by using the equations with correlation that Tony provided in lecture and have Excel solve a system of equations to find the b coefficients. If you need more detail on how these equations were derived please check the lecture notes, book or ask for help.

To begin with we need to pick out what variables we would like to use in our model. For this example we will pick a "y" variable to predict and two "x" variables (also called predictors).

Tony showed that we can find the coefficients corresponding to a regression equation in dimensionless "z" units (units of standard deviations) which for two predictors looks like:

$$z_y = b_{1.2} * z_{x_1} + b_{2.1} * z_{x_2} \qquad (1)$$

We can solve for the "b" coefficients in z units using the correlations between y, x1 and x2 as follows:

$$b_{1.2} + b_{2.1} * corr_{X_1X_2} = corr_{YX_1}$$
$$b_{1.2} * corr_{X_1X_2} + b_{2.1} = corr_{YX_2} \qquad (2,3)$$

To solve the above equation, you first try to reduce the two variables ($b_{1.2}$ *and* $b_{2.1}$) to one variable, say, $b_{1.2}$ by dividing the first equation by $corr_{X_1X_2}$ for each term,

$$\frac{b_{1.2}}{corr_{X_1X_2}} + b_{2.1} = \frac{corr_{YX_1}}{corr_{X_1X_2}} \qquad (2')$$

Subtract equation (3) from equation (2') to get one equation for one variable,

$$b_{1.2}\left(\frac{1}{corr_{X_1X_2}} - corr_{X_1X_2}\right) = \frac{corr_{YX_1}}{corr_{X_1X_2}} - corr_{YX_2} \qquad (3')$$

Solve for $b_{1.2}$ from (3') to get the following:

$$b_{1.2} = \frac{corr_{YX_1} - corr_{YX_2} corr_{X_1X_2}}{1 - \left(corr_{X_1X_2}\right)^2} \qquad (2'')$$

Substitute (2") back to (2) to get the coefficient $b_{2.1}$

$$b_{2.1} = \frac{corr_{YX_2} - corr_{YX_1} corr_{X_1X_2}}{1-\left(corr_{X_1X_2}\right)^2} \quad (3")$$

We can also have excel do it for us by using Matrix math. Consider b1.2 and b2.1 as the variables you are trying to solve for and write an array of the coefficients of these variables and the right hand side of the equation. This will look like:

1	Corrx1x2	Corryx1
Corrx1x2	1	Corryx2

If we consider the coefficients as matrix A and the right hand side of the equations as matrix C then we can solve for the b's as:

$$\begin{bmatrix} 1 & corr_{x_1x_2} \\ corr_{x_1x_2} & 1 \end{bmatrix} \begin{bmatrix} b_{1.2} \\ b_{2.1} \end{bmatrix} = \begin{bmatrix} corr_{yx_1} \\ corr_{yx_2} \end{bmatrix} \Rightarrow A*B=C \quad (4,5,6)$$
$$B = inv(A)*C$$

To have excel calculate B for you use the matrix inverse function (minverse) and the matrix multiply function (mmult). These functions are array functions which means you need to highlight an array (2 cells in a column) for output before solving. **both functions must be executed with command-enter **

These regression coefficients were derived for a function of dimensionless z's, so we want to convert them back to original units. This can be done by substituting the appropriate equations from (7) to convert equation 1 to a form involving y, x1 and x2.

$$\begin{aligned} y &= \bar{y} + z_y SD_y \Rightarrow z_y = (y-\bar{y})/SD_y \\ x_1 &= \bar{x}_1 + z_{x_1} SD_{x_1} \Rightarrow z_{x_1} = (x_1 - \bar{x}_1)/SD_{x_1} \\ x_2 &= \bar{x}_2 + z_{x_2} SD_{x_2} \Rightarrow z_{x_2} = (x_2 - \bar{x}_2)/SD_{x_2} \end{aligned} \quad (7)$$

Now we have a predictive equation that reads:

$$y = coeff_1 * x_1 + coeff_2 * x_2 + constant$$

We need to asses the success of our regression, so lets calculate an R for the prediction. Recall from lecture that:

$$R = \sqrt{b_{1.2} * corr_{YX_1} + b_{2.1} * corr_{YX_2}} \quad (8)$$

and that the standard error can be calculated by:

$$SE_{zerocorr} = \frac{1}{\sqrt{n-k}} \quad (9)$$

We can test R by dividing R/SE and testing against some Zcritical. It follows that we can also calculate the minimum R needed to be significant by multiplying Zcritical*SE.

Project 1: Asses the effect of adding a second variable to a regression for temperature in Yuma, Arizona.

Many stations in the southwestern U.S. have had a marked warming trend from the 1950s to the present. Yuma is a good guinea pig station because it is not in a large metro area (like Phoenix is) so that urban growth can be ruled out. Arizona temperature may not be much affected by ENSO, but lets explore. The season during which gradual warming is most noticeable in the southwestern U.S. is Jan-Feb-Mar, so using an average of the temperature for those three months for each year would be best. Calculate the correlation between temperature (averaged fro JFM) in Yuma, Arizona and CO_2 measurements from Mauna Loa, Hawaii. Is the correlation statistically significant (recall that this is the same as asking if the slope is significantly different than zero)?

Next we will perform a multiple regression of temperature in Yuma against Mauna Loa CO_2 as x1 and Nino 3.4, an El Nino index, as x2. Calculate the coefficients, correlation and regression equation. Is the correlation significant? Did the correlation change much by adding a second predictive variable? What does this tell us?

Project 2: Tropical Pacific temperature signal
Tarawa is an island in the tropical Pacific, near the equator and near the international dateline. It would be expected to be affected both by gradual warming (eastward expansion of the western tropical Pacific warm "pool" SST) and by ENSO (it becomes warmer when there is an El Nino, cooler when there is a La Nina). The ENSO effect would be strongest during the season having mature El Ninos/La Ninas, which is October through December. An average of each year's Oct-Nov-Dec temperature would be the ideal variable, since single month temperatures are "noisier".

Use the OND averaged temperatures to perform a multiple regression against Nino 3.4 and Mauna Loa CO_2. Was the regression successful? What does the outcome tell you about the data? Can you justify including both variables? Why or why not? Do you think this is a good model to use to predict temperature at Tarawa?