5 Evaluating Correlations

a. Significance test of difference between two sample composite means

A common compositing approach is to contrast circulation features associated with two extremes of an index: wet (dry) years in Utah precipitation or El Nino/La Nina seasons. Sample means can be computed from the same population or completely different batches of data. An appropriate null hypothesis is that the population means are the same. Either a 1-tailed or 2-tailed test can be done. Without proof (see Wilks), we can derive an appropriate t –test value as:

- $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{((n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2)(1/n_1 + 1/n_2)/(n_1 + n_2 - 2)}$
- where $\bar{x}_1$ and $\bar{x}_2$ are the two sample means, n_1 and n_2 are the size of the samples, and s_1 and s_2 are the sample standard deviations. The number of degrees of freedom is $n_1 + n_2 - 2$ to reflect that both sample means are specified.
- The matlab command ***ttest2*** can be used only in the situation where the two sample standard deviations are the same (so don't rely on it).

b. ANOVA and significance test of linear regression

The focus shifts now to assessing how much confidence we can have in linear regression. Common practice is to assume that when the linear correlation between two variates exceeds 0.5 or 0.6, then we have at least a practical and useful association between the two variables, if the linear correlation is determined from a large sample. This sort of threshold is of use because the explained (unexplained) variance is at least 25-36% (75-64%) of the total variance. For environmental fields with a large number of degrees of freedom, it is possible to have a linear correlation between two variates be as low as 0.1 and still potentially be judged to be statistically significant. Such low correlation values may not have any practical significance to estimate one variable from the other. However, they may help point out a physical relationship between the two variables that was unknown before. The data may then be transformed, filtered or combined with other data to develop some predictive relationship.

Begin by reviewing how linear regression was determined by estimating the best linear fit of y given the variable x (Chapter 4). Define the line of best fit by $\hat{y}_i{}' = \mathrm{b}\, x_i{}'$ or $\hat{y}_i - \bar{y} = \mathrm{b}\,(x_i - \bar{x})$. Rearranging, then $\hat{y}_i = \mathrm{a} + \mathrm{b}\, x_i$ where a = $\bar{y} - b\bar{x}$. Also, $s_y{}^2 = b^2 s_x{}^2 + \overline{e_i^2}$, which states that the total variability of y is given by the sum of the variability explained by the variable x plus the error variance. In terms of the percent of the total variability of y, we can write this equation as 1 (100%)= r^2 (percentage of explained variance) + [1- r^2] (percentage unexplained or the error variance). Traditionally, this relationship is written in a slightly different form, i.e., the sum of squares SS form $ns_y{}^2 = nb^2 s_x{}^2 + n\overline{e_i^2}$.

ANOVA Table- Regression Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST=n s_y^2	n-1	n s_y^2 /(n-1)	
Regression	SSR=n $b^2 s_x{}^2$	1	MSR=n $b^2 s_x{}^2$ /1	(n-2) $b^2 s_x{}^2$ /($s_y{}^2 - b^2 s_x{}^2$)
Error	SSE=n ($s_y{}^2 - b^2 s_x{}^2$)	n-2	MSE= n ($s_y{}^2 - b^2 s_x{}^2$)/ (n-2)	

All of this information is summarized in analysis of variance (ANOVA) tables, such as that shown above. ANOVA is a common way to summarize whether the variance of variable y explained by variable x is large in terms of three measures: mean squared error of the regression (MSE), variance explained by the regression (MSR), and the F ratio that is assumed to have a known parametric form (Fisher's F distribution). We want: (1) the scatter around the line of best fit to be small, i.e., that SSE and MSE are small; (2) the percent variance explained by the regression to be large (or MSR large); and (3) that the F ratio is large, which is the ratio of the explained variance to that of the error. Note that two degrees of freedom are used to specify the regression line (the coefficients a and b). If there are only two observations, then the regression line would pass through the two points and there would be no error (MSE=0). Another way of thinking about it is 1 degree of freedom from the total n is used by the mean value of y and the other is used by the regression coefficient. When looking at some statistical sources, be aware that the MSR should always be greater than the MSE unless n is small and the linear correlation is small as well. You will see that the parametric distribution of F is determined entirely by the degrees of freedom of the larger MS (the regression) and the degrees of freedom of the smaller MS (the error).

The following correlation form of the ANOVA table emphasizes the key values: the correlation coefficient (which provides the explained variance and the unexplained variance) and the degrees of freedom of the regression (1) and the degrees of freedom of the error (n-2).

ANOVA Table- Correlation Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST= n	n-1	n/(n-1)	
Regression	SSR=n r^2	1	MSR=n r^2 /1	(n-2)r^2 / $(1-r^2)$
Error	SSE=n (1- r^2)	n-2	MSE= n (1-r^2)/(n-2)	

So, let's compare two climate indices in January: (1) the MEI index that describes variability in the tropics and (2) the PNA index that describes variations over North America. These indices are available from: http://www.esrl.noaa.gov/psd/data/climateindices/list/. Their time series don't look too similar (Fig. 5.1) and when their linear correlation is computed r = 0.32, which means it's going to be hard to pass a practical sniff test, these two variables share ~10% of their variance in common. In this instance, F = 7.0. We compute the probability, f, that we can reject the null hypothesis (that the correlation observed between these two variables could happen by chance) assuming a 1% risk according to the Fisher F parametric distribution using the Matlab

finv function and the appropriate degrees of freedom. The matlab command **f = finv(.99,1,68)** finds the value that 99% of the samples from a F parametric distribution should be less than for the specified numbers of degrees of freedom. In this case, f = 7.0 < F= 7.3, so we would reject the null hypothesis in this instance. There is a 1% risk that the variability shared in common by these two indices could have simply happened by chance. But, wait! Does that make sense? Are we being too easy? Do these two indices really look that close to one another? What happens if we accept a 5% risk to falsely reject the null hypothesis? Ah, then we only need to beat f=4. Here's the slippery slope of evaluating the significance of results- are you going to shop for the test and thresholds that prove your point? First, do we really have 68 independent values? What if we were really conservative and only assumed half the years as being independent? **f = finv(.99,1,34)= 7.4.** Then, we would have to reject the null hypothesis. Hmmn, in other words, it is pretty easy to reject the null hypothesis often when the results are pretty tenuous.

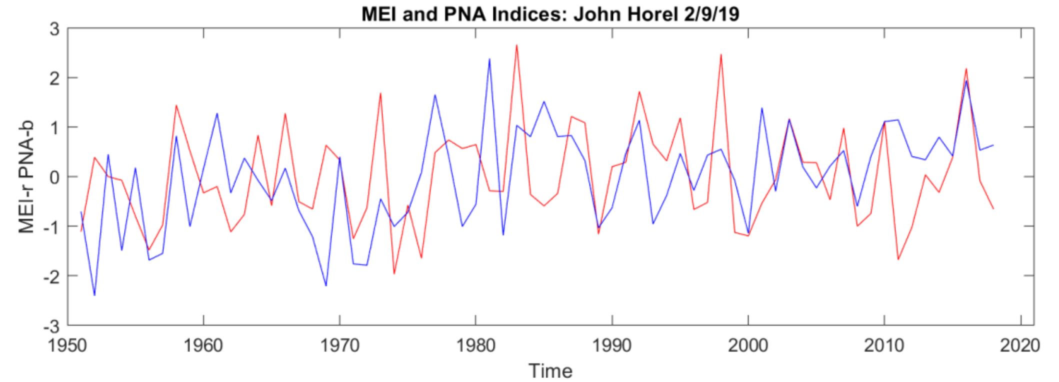

Figure 5.1. Time series of MEI and PNA indices for January only.

ANOVA Table- MEI/PNA

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST= 68	67	1.0	
Regression	SSR=6.8	1	6.8	7.3
Error	SSE=61.2	66	0.9	

c. Bootstrapping

The methods described so far to assess whether a mean value or linear correlation differs substantively from the hypothesized value assume some parametric assumptions such as that the population is distributed according to the Gaussian distribution. Many nonparametric approaches are available as well. The first one is bootstrapping, where the original sample is resampled many times. As an example, let's return to the linear correlation coefficient between the MEI and PNA indices that was found to be 0.32. This was based on 68 monthly values. What if we resampled these two data sets and removed one pair of values and recomputed the linear correlation? Then, we started over again and randomly removed another pair of values and continued 10,000 times. This approach is called resampling with replacement that leads to the histogram in Fig. 5.2 of the 10,000 linear correlation coefficients. Removing only a single monthly value and recomputing the linear correlation each time yields quite a bit of spread (standard deviation of 0.11), which means there is a ~95% chance you could get values anywhere between 0.10 and 0.54 (~±2 standard deviations from the mean). That doesn't give me a lot of confidence either about this relationship.

d. Cross validation

A traditional approach for assessing the confidence of linear regression results obtained from a sample is to apply the linear regression to an independent sample. The data could be divided in half at the outset and only half is used to "train" the regression and the rest is kept for the verification of the regression. How the data are split between the "dependent" and "independent" samples can be tricky, especially if there are long term trends or other systematic behavior within the data set.

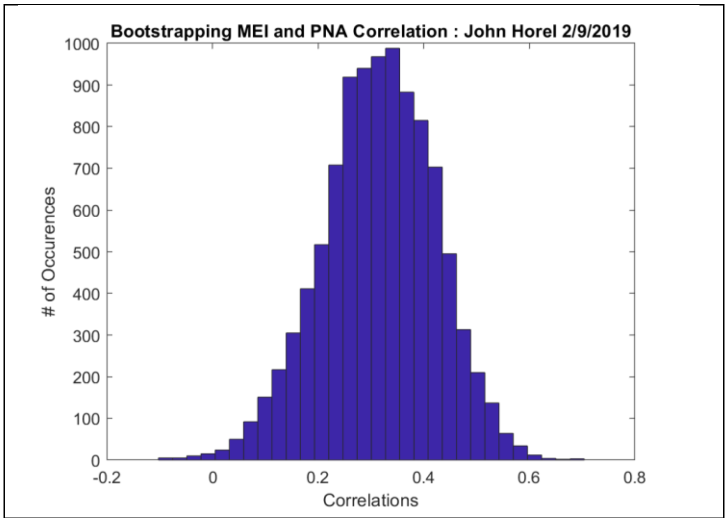

Figure 5.2. Bootstrapping of MEI-PNA correlation coefficients.

A more rigorous cross validation approach is to systematically remove data elements. One method is to remove 1 value, compute the linear regression from the n-1 values, then evaluate how close the linear regression estimate is to the withheld value. This can be repeated over all n possible values. There's nothing particularly special about removing only 1 value- you can withhold more. When a field has a lot of serial dependence, then it becomes particularly important to withhold enough observations to actually affect the linear regression (i.e., if the number of degrees of freedom in the data set is much less than the total number, then you need to withhold a larger number of values).

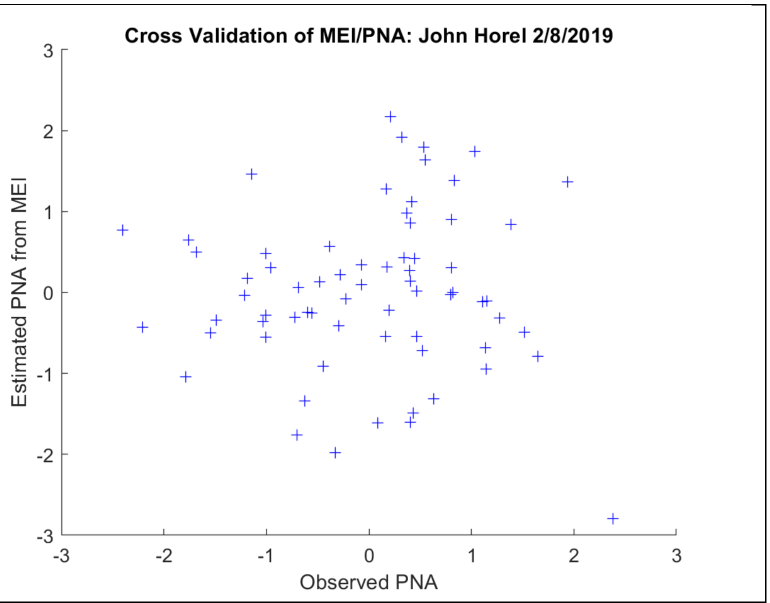

Figure 5.3. Observed PNA values vs. those estimated from the MEI index obtained by cross validation.

The code includes a cross-validation of the January MEI-PNA index values. The results of this analysis are shown in Fig. 5.3. The amount of scatter is very high because remember that the linear correlation is 0.32, i.e., the explained variance is order 10%. Note in the code that our estimate of the PNA index value in any particular year is derived from the regression estimate (small explained variance) and a random element added (with large unexplained variance in this situation). The mean difference (bias) between the values estimated and those observed are close to zero, which implies that there is no systematic bias in estimating the PNA index value from the SOI value. However, the root-mean-squared error is large (1.4), which provides a useful measure of how confident we can be in using this linear estimation. Looking at Figure 5.3, we shouldn’t have a lot of confidence that we can estimate the PNA state from the MEI index.

Now there are good physical reasons for why conditions in the tropical Pacific are related to mid-latitude circulation features. But taking two indices and blindly comparing them is not the way to make that case. Do the sea surface temperature conditions precede the midlatitude fluctuations? Should we look at more than January, maybe the entire winter season? The main message is to think about the hypothesis carefully before jumping into statistical analysis.