1.5 Statistical significance of linear trend

Using the least square method, linear trends (regression coefficient) can be calculated for any time varying data. When we perform the linear regression, we often find non-zero trend (slope) but how can we tell whether or not the trend is statistically meaningful? For example, we found a positive regression coefficient in Figure 1.12. One way to measure the importance of the trend, we calculated the R^2 value which measures the fraction of variance explained by the trend.

We can also perform a hypothesis testing to assess the significance of the trends. Here we ask the following hypothesis. We state the null hypothesis and its alternative with the 95% confidence level.

H0: There is no significant trend.

H1: There is a significant trend (the regression coefficient is significantly different from zero).

To rephrase **H0**, the observed regression coefficient a_1 is not different from zero given the uncertainty in estimating the trend, and we use the 95% confidence level as the threshold. Note that this hypothesis is formulated for a two tail test because we are not specifying whether the trend would be positive or negative. Alternatively, we can formulate the one tail test as follows.

H0: There is no significant positive trend.

H1: There is a significant positive trend.

This set of hypothesis is not considering the possibility that the trend can be negative. Our calculation from the previous section indeed shows the positive trend, so it is acceptable to perform the one tail test. One tail test gives more power to reject the null hypothesis because it slightly lowers the boundary of the confidence interval.

The residuals of regression, e, are defined as $e = (a_0 + a_1 x) - y$, and its variance, s_e^2, is

$$s_e^2 = \frac{1}{N_{eff} - 2} \sum_i^N e_i^2 \tag{1.33}$$

This measures the scatter of data about the regression line. N_{eff} is the effective sample size. *If the values of e_i are independent*, we have $N_{eff} = N$, the standard error of the regression coefficient, and s_a, is shown to follow (Santer et al., 2000)

$$s_a^2 = \frac{s_e^2}{\sum_i^N (x - \overline{x})^2}. \tag{1.34}$$

The statistic, $t = a_1/s_a$, is distributed as Student's t with the degree of freedom of $N_{eff} - 2$.

In this example, there are many years of observation. For the whole period (1879 to 2015) the linear trend a_1 is 0.0125°F/yr and s_a is $0.004° F/yr$. Then the t-statistic is

$$t = \frac{a_1}{s_a} = 3.05. \tag{1.35}$$

The critical t value for the two-tail test with $d.f. = N_{eff} - 2$ and 95% confidence level is 2.0. The t-value of 3.05 is outside the envelope of uncertainty about the zero trend.

Thus, we reject the null hypothesis, and we conclude that there is a significant century-scale trend (1879 to 2015). Using the standard error of the linear trend, we may also state that the linear trend and its uncertainty (95% CI) is $1.25 \pm 0.8 \times 10^{-2\circ}$F/yr where the uncertainty is two times the standard error. Here, we performed the two tail test. We can repeat the analysis with one tail test and we know that the null hypothesis will be rejected because it didn't pass the two-tail test.

Effective sample size

In the previous example, there was a caveat that each measurement may not be independent from one another. Then the degree of freedom is modified based on the effective sample size, N_{eff}. In our example, we were looking at the temporal evolution of temperature. It is possible that measurement from one year is NOT independent from prior year or the year after. If the temperature in one year is similar to the previous year's temperature, we may have an issue with the independence of the data. There is a method to calculate the 'effective' sample size including the effect of persistence in the data (Breatherton, 1999).

$$N_{eff} = N \left(\frac{1-r}{1+r} \right) \qquad (1.36)$$

where r is the lag-1 autocorrelation coefficient. Autocorrelation is a correlation with itself but with some time lag. Without the time lag, the correlation must be perfect (=1). Lag-1 autocorrelation means that you take the correlation coefficient with the same variable itself but it is shifted in time by 1 unit. If the signal is persistent in the data, you may see non-zero autocorrelation ($r > 0$), and in such case, we scale down the effective sample size. It is always a good idea to use this method for determining the effective sample size of the observational data.

Statistical significance of correlation

We can also test whether or not the two variables are correlated in a significant way. Let's think about this example. We have two variables, surface air temperature and an index for El-Nino condition, and we want to investigate the relationship between air temperature and the state of El-Nino condition.

We go ahead and calculate the correlation coefficient between the surface air temperature and the El-Nino index for each grid cells according to Eq 1.27. Following our thinking about the hypothesis testing, we state the null hypothesis and its alternative.

H0: There is no significant correlation.

H1: There is a significant correlation (the correlation coefficient is significantly different from zero).

In this case, the t-statistic is

$$t = r\sqrt{\frac{N-2}{1-r^2}}. \tag{1.37}$$

And the effective sample size depends on the lag-1 autocorrelation of the two variables, surface air temperature (r_1) and the index of El-Nino condition (r_2).

$$N_{eff} = N\left(\frac{1-r_1r_2}{1+r_1r_2}\right) \tag{1.38}$$

Exercises We will test the significance of the linear trend in the monthly surface air temperature data from NCEP reanalysis (air.mon.mean.nc).

1. Calculate the linear trend of January temperature for each grid cell. (Store the results in an array.)

2. Calculate the standard error of the linear trend for each grid cell.

3. Calculate the effective sample size for each grid cell.

4. Perform a t-test to determine whether or not there is a statistically significant linear trend for each grid cell.

5. Make a color map (using m_pcolor) of January temperature trend. Place a marker for the grid cells with statistically significant trend.

6. Publish the MATLAB script that performs all of the activity above, and submit it as a report in the PDF format.

Reference

Santer, B.D. et al., (2000) Statistical significant of trends and trend differences in layer-average atmospheric temperature time series, *Journal of Geophysical Research*, (105), D6, pp. 7337-7356.

1.6 Correlation maps

Correlation analysis can be a very powerful tool to establish a statistical relationship between the two variables. Section 1.4 showed that a correlation coefficient between a pair of two variables x and y is defined as the covariance divided by the standard deviation of x and y (Eq. 1.27). Then section 1.5 further developed the method to evaluate the statistical significance of the correlation coefficient.

Consider the example of surface air temperature and the index for El-Nino. The surface air temperature has the spatial and temporal variation, $T = T(t, x, y)$, where x is longitude, y is latitude and t is time. Then it is possible to construct a $x - y$ map of correlation coefficient by performing the calculation of the correlation coefficient many times, for each position in (x, y) space. Repetitive calculations can be done by the loop statement in MATLAB. For example, you can define a new function ”correlate” that calculates the correlation coefficient and its statistical significance, where $m(t)$ is the index of El-Nino and $T(t, x, y)$ is the surface air temperature.

```
>> for i=1:Nx
>>    for j=1:Ny
>>       [r,s] = correlate(m,T(:,i,j));
>>       rxy(i,j)=r;
>>       sxy(i,j)=s;
>>     end
>>  end
```

Here, we have to be careful that both the indices of El-Nino and temperature must be anomalies from the climatology. It usually means that the mean seasonal cycle is subtracted from the original data. El-Nino events have strong interannual variability (2–7 year timescale) and this step allows us to focus on the year-to-year variability. The raw data often includes strong signature of seasonality, and it should be removed before calculating the correlation. Fig 1.13 shows the resulting pattern of rxy.

Exercises We will generate correlation maps between surface air temperature and the index of North Pacific Gyre Oscillation `http://www.o3d.org/npgo/`.

1. Develop a MATLAB function that calculates the correlation coefficient between two variables.

2. Download the NPGO index from Prof. Di Lorenzo's website (`http://www.o3d.org/npgo/npgo.php`) and load it on MATLAB.

3. Remove the mean seasonal cycle from the data and calculate the effective sample size of the air temperature data for each grid cell for the 66 year period (1950-2015).

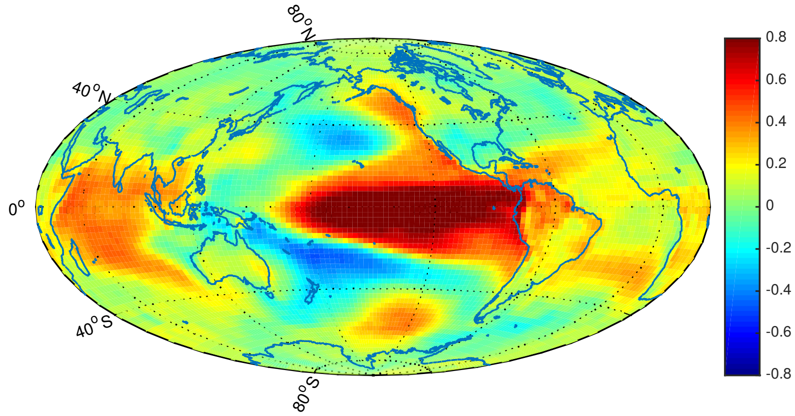

Figure 1.13: Correlation map. Surface air temperature anomalies are correlated with the Nino 3.4 index. The data period is from 1950 to 2015.

4. Calculate the effective sample size of the NPGO index for the 66 year period (1950-2015).

5. Calculate correlation coefficient for each grid cell.

6. Generate a map similar to Fig 1.13 for the NPGO.

7. Publish the MATLAB script that performs all of the activity above, and submit it as a report in the PDF format.

Reference

Di Lorenzo et al., 2008, North Pacific Gyre Oscillation links ocean climate and ecosystem change, Geophysical Research Letters.