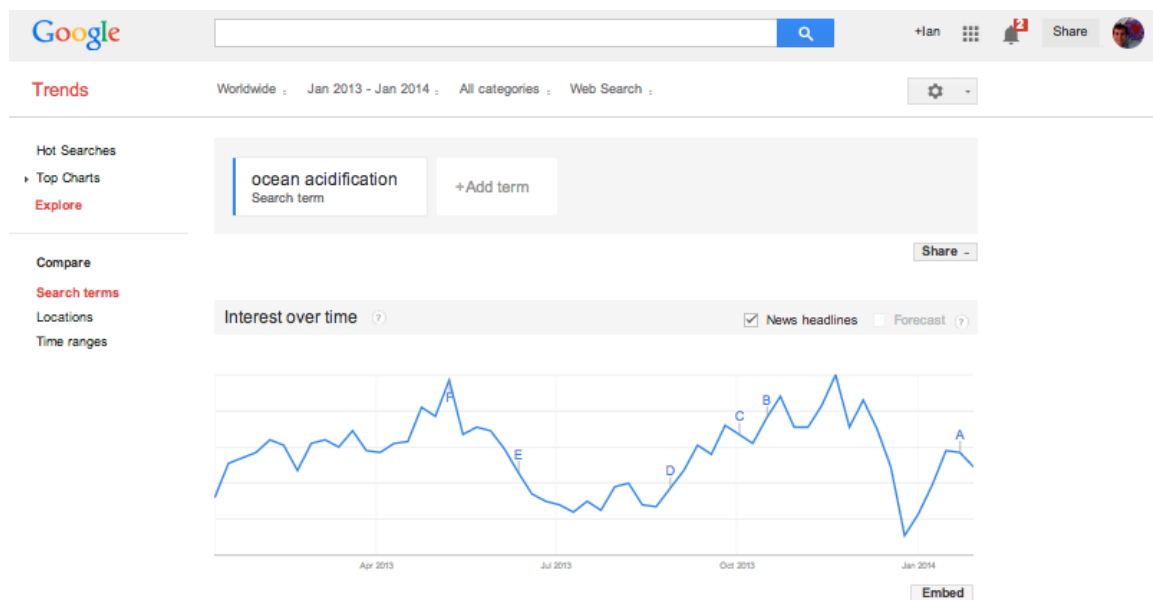


Regression and Time Trends Exercise: Measuring temporal autocorrelation using the auto-correlation function (ACF).

For this exercise, you will need Microsoft Excel, or another spreadsheet program that can read compatible files, along with the Excel Workbook available at (https://dl.dropboxusercontent.com/u/596355/ocean_acid_autocorr.xlsx).

This spreadsheet contains data from Google Trends (<http://www.google.com/trends>). Google Trends tracks the frequency of web searches through the Google search engine, and has been used as an index of how many internet-connected users are interested in a particular topic over time. The data we will work with represents the relative number of searches for the phrase “ocean acidification” for each week from January 2013 to January 2014.



In the last exercise, we explored why temporal autocorrelation can prevent us from effectively using linear regression to judge whether something is trending over time. Recall that what autocorrelation does is give us false-confidence in the value of the slope of the long-term trend over time, but how big a problem is it?


To figure this out, we need some measure of the strength of autocorrelation in a given series of measurements. We typically do that by calculating a statistic called the autocorrelation function (ACF). This is a correlation coefficient similar to the Pearson Correlation, which measures the strength of association between two different variables. The key difference here is that instead of two different variables, we are measuring the association between measurements of a system at some time k to measurements of that same system a certain number of observations into the past. The number of measurements backwards-in-time over which we are measuring those differences is called the lag, h .

The formula for the autocorrelation function at a given lag, h , is given by:

$$\hat{p}(h) = \frac{\sum_{k=h}^T (y_k - \bar{y})(y_{k-h} - \bar{y})}{\sum_{k=h}^T (y_k - \bar{y})^2}$$

This equation compares measurements at time k , (y_k) and measurements at the previous times (y_{k-h}) to the mean of all observations in the series (\bar{y}), for all times for which we have measurements. The resulting number varies between -1 and 1. Values close to 1 indicate that observations separated by a lag of time h are very similar to each other (positive autocorrelation), while values near -1 indicate that they are very different (negative autocorrelation).

Lets apply this equation to some real data.

- Open the Excel spreadsheet “ocean_acid_autocorr.xlsx”. The first two columns represent the date and the normalized search volume (our measurement of how popular the term “ocean acidification” was on Google). The next column is the same data but shifted “down” by one cell. This means that every row in the table now has the search volume at a particular week in Column B, and the search volume in the previous week in Column C.
- The next two columns are different parts of the equation for the autocorrelation function. Double-click on cell D4 to see the formula in that cell. It is using the formula AVERAGE() to find the mean value for all of the cells in that particular column, and is using that number to come up with the numerator component of the ACF for each time-interval. Cell E4 contains the denominator part of the equation.
- In Columns C and D, drag the fill box to fill in these equations for the blank cells in the rest of the table (hint, highlight cells D4 and E4, then click on the box in the lower-right corner  and drag the box down).
- Formulas are already set up to put together the parts of the ACF equation. Cells D56 and D57 sum up the values in Columns D and E (the $\sum_{k=h}^T$) parts, and Cell B57 takes their ratio.
- That’s it! We’ve just calculated the ACF at a lag of one observation, which compares values that were measured consecutively in time.

Use the spreadsheet to answer the following questions:

1. Examine the graph of the time-series at the top of the worksheet. Is there any visual evidence of autocorrelation in the data? If so, what is it?

2. What is the value of the ACF that you calculated? Does this confirm or deny your suspicion about autocorrelation?
3. In this analysis, we calculated the ACF for consecutive observations (a lag of 1). How would you need to modify the spreadsheet in order to calculate the ACF for observations that are separated by two time-steps (a lag of 2)?
4. Perform the modifications that you proposed in the previous question. What is the lag 2 autocorrelation? How is it different from the correlation at a lag of one?