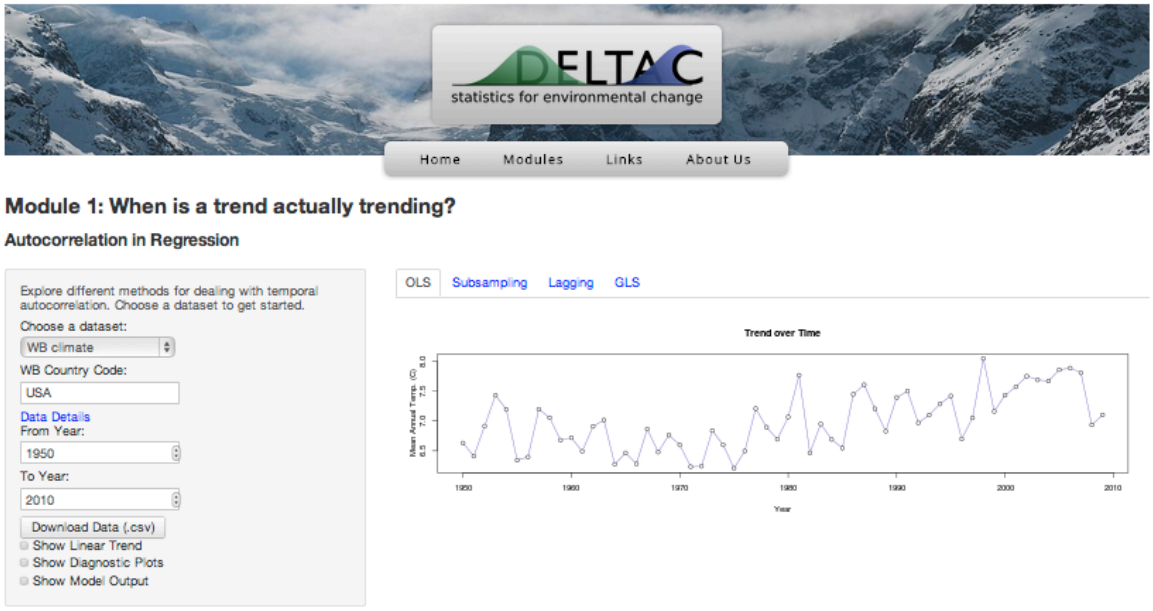
Regression and Time Trends Exercise: Plotting the ACF.

In the last exercise, we explored how to measure the strength of temporal autocorrelation using the autocorrelation function (ACF) and assess whether it might be a problem for making inference about trends over time using linear regression. The calculation of the ACF itself is labor-intensive, so researchers typically employ software to help us assess autocorrelation in the data over many different lags.

In this exercise, we will use an interactive web application to examine patterns of autocorrelation in time-series data. You can access the application at: (http://spark.rstudio.com/statmos/mod1_regression). You should see a web page that looks like this:



This application allows you to search online databases for time-series data, plot the data, apply linear regression, and view diagnostic plots to help see if you can make valid inference using Ordinary Least Squares. If not, it gives you a few alternative strategies for assessing whether something is trending over time. For now, we will focus on interpreting diagnostic plots that help us check whether a regression analysis is valid for time-series data. Follow the instructions below, and then answer the questions at the end of the handout.

Instructions:

Spend a few minutes exploring the different options that the application gives you, but stick to the tab that is labeled 'OLS' for now. Here's a brief field-guide to some of the options.

Choose a dataset:

These options, and the ones below, allow you to select the data you want to plot and examine. Two types of data are available:

- The dataset "WB climate" contains temperature trend data over the past century summarized by country. You can search for a particular country by typing in a three-letter code that corresponds to the country name (you can find the list of codes at (http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3)
- The dataset "Google ngrams" contains data on the frequency that words and phrases appear in all of the books indexed by Google Books, currently numbering approximately 6-million. You can search for a particular word or phrase by typing it into the box.

From Year and To Year

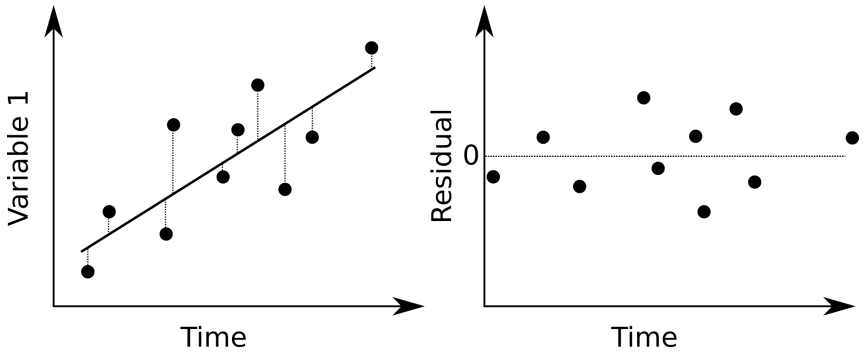
These options control the time-range of the data that is plotted.

Show Linear Trend

Checking this option will display a best-fit regression line over the data plotted in the right-hand panel.

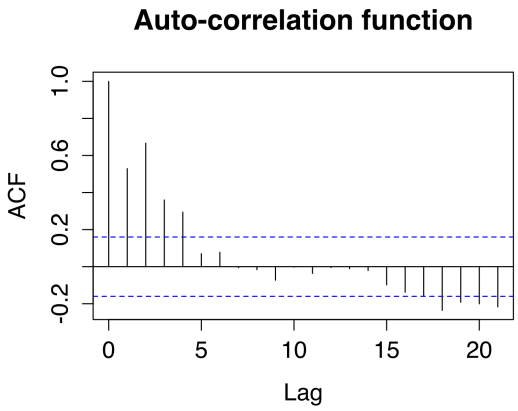
Show Diagnostic Plots

The first of the two diagnostic plots show the residual variation in the data that is not "explained" by the regression model. Essentially, these are the differences between the observed data, and the value that would be predicted using the model.



Remember that a major assumption of linear regression is that the residuals be independent of each other, normally distributed, and have a constant variance over the range of the data. This plot allows you to assess whether those assumptions are reasonable for the data at hand.

The second plot shows the ACF for the model residuals at many different values of the lag. In the last exercise, we calculated the ACF for observations that were adjacent in time. This plot shows that value, along with the values of the ACF at lags greater than 1.



In this plot, the height of the vertical bars shows the strength of autocorrelation at each lag. Because observations are always perfectly correlated with themselves, the ACF at a lag 0 is always equal to one. ACF values typically decrease as we examine observations separated by larger numbers of observations, like in the plot above. The horizontal dotted lines indicate the maximum amount of autocorrelation that we could expect by chance. If the bars are taller than the dotted line, then it indicates that the residuals are not independent, and we should be suspicious of the results from our regression model.

Show Model Output

Checking this option will display a standard output from the linear regression model fit to the data.

Questions:

1. Refresh the page in your web browser to return to the default options, then check the "Show Linear Trend" and "Show Diagnostic Plots" options. This data shows mean annual temperature in the USA from 1900 to 2010. Does there appear to be an increasing or decreasing trend in temperatures over time?

2. Examine the diagnostic plots. Do the residuals from the regression model appear to meet the assumptions that are required for valid inference using linear regression? If not, which assumptions appear to be violated, and how?

3. Change the "From Year" field from 1900 to 1960. Examine the diagnostic plots. Do the residuals from the regression model appear to meet the assumptions that are required for valid inference? If not, which assumptions appear to be violated, and how?

4. Change the "From Year" field to 1930 and the "To Year" field to 1980. Examine the diagnostic plots. Do the residuals from the regression model appear to meet the assumptions that are required for valid inference? If not, which assumptions appear to be violated, and how?

5. How did the trends differ in the two time-periods you examined in questions 3 and 4 (1960-2010 and 1930-1980)?

6. How are the trends you examined in Questions 3 and 4 related to potential problems with using linear regression to assess whether there is a long-term trend in temperatures from 1900-2010?