Lab 3: Variance, Correlation, and Covariance in the UBC Weather Station Data

Lab Overview

In this lab you will use linear regression, correlations and multi-linear regression to ask the following three questions for the UBC weather station/global temperature time series:
-Is there a trend in the temperature time series?
-Are temperature time series correlated with external or internal factors (e.g. atmospheric CO2, El Nino oscillation)?
-Are your results compatible or not with the IPCC reports, or can they prove or contradict that there is a global warming trend and that it is caused by anthropogenic CO2 emissions?

Learning Goals

After this lab you should be able to:

- Perform linear regression and multilinear regression, calculate correlation coefficients, and use histograms in Matlab
- Use these tools to quantify the importance of different factors in explaining the variability of a time serie
- Think critically about your plots/results, and assess whether they allow you to make conclusions on a hypothesis or not.

To hand in

1. A figure containing the raw time series of the 7 variables that will be used in this lab, in separate subplots, with linear regression for temperature time series (part 1)

2. One figure containing histograms of temperature before/after 1985 for local (UBC) and global temperatures (part 1)
3. One figure containing a scatterplot of the UBC temperature against the global mean temperature (part 3)
4. One figure containing the scatterplots of UBC temperature against MEI index and global temperature against MEI index, and associated linear regressions (part 4)
5. One figure containing the scatterplots of global mean temperature anomaly against temperature predicted by a multilinear regression (using 5 explanatory variables, part 5)
6. Two figures containing, respectively, global temperature anomaly as a function of Donald Trump′s age, and made-up global temperature against CO2 time series (part 6)
7. The Lab 3 Worksheet

1. Load the data and plot each time series

Start by loading the data needed for this lab. They are all contained in the file 'lab3_data.xlsx' available on connect. You can open it with the xlsread() function which we used in previous labs. The spreadsheet contains the following 8 variables:

- **The date**, which is the same for all of the time series and has been formatted nicely as a number corresponding to the year+month/12. For example, June 2000 would be "2000.46". Thus**, you will not need to bother with converting dates from text to datenumber or datevec** for this lab! All time series in this lab have the same starting/end date and are monthly values, so **you only have one vector of date to deal with**!

- **Two temperature anomalies time series**. The first one is measured at UBC weather station; you computed the anomaly last lab. The second one is the global mean temperature collected by the Climate Research Unit. Both time series are anomalies relative to the 1961-1990 seasonal cycle and are expressed in Celsius. The UBC time series contains NaNs at the beginning because measurements only began in ~1960

- **The total solar irradiance (TSI) time series**, i.e., the flux of solar energy entering through the top of the atmosphere in W/m^2. The datasets from the SATIRE project (monthly mean) and from Lean et al (2000) (yearly mean) have been combined to obtain monthly TSI over 1950-2016.

- **The global mean stratospheric aerosol optical depth (AOD)** time series (dimensionless) obtained from the Goddard Institute for Space Study, which mostly reflects long-lived aerosols from volcanic eruptions. A constant value was assumed for the period 2012-2016 for which no data was available.

- **The atmospheric CO2 concentration and anthropogenic SO2 emissions:** yearly time series were obtained from the Earth Policy Institute and the Pacific Northwest National Laboratory, respectively, and interpolated to obtain monthly time series. Their units are ppm and Tg/year (Teragrams/year, i.e. 10^9 kilograms/year) respectively

- **The Multivariate ENSO Index (MEI)** which we used in lab 1 and was reviewed during the prelab. The MEI is dimensionless and positive during an El Nino (negative during La Nina) event.

Plot each time series in separate subplots but on the same figure using the function subplot() (seen in several lab already). I would recommend using a 4x2 subplot configuration. Make sure to properly label your plots.

In addition and as a warm-up, perform a linear regression of each temperature time series against time using the function regress() (cf prelab):

- Display the corresponding regression line on each graph. The regression line is the temperature predicted by your regression as a function of your independent variable (i.e.

the date in this case)
- The slope of each regression will give you the temperature trend.
- Display the values of the trend (don't forget units!) and the coefficient of determination on each temperature time series plot either in the legend (use the function legend()) or using the function text().

Are you surprised by the sign of the trend found? Looking at 95% confidence intervals on these slopes, are you confident that the trend observed are significant? Which temperature time series has the highest trend? The highest coefficient of determination? Does this surprise you?

Last, we will use histograms to visualize temporal trends in temperature time series. If there is a monotonous trend in our data, the distribution of temperature anomalies before, say, 1985 and after 1985 should be shifted. For each temperature time series, plot these two histograms on the same plot. You can make a single figure with two subplots (one for UBC data and one for the global data). To plot the two histograms in each subplot, use command lines of the form:

```
histogram(T(mask1),linspace(min(T),max(T),n),'Normalization','probability')
hold on
histogram(T(mask2),linspace(min(T),max(T),n),'Normalization','probability')
```

where T is your temperature time series, and mask1/mask2 are masks to select data before or after 1985. `linspace(min(T),max(T),n)` set the edges of the bins of your histograms to be regularly spaced between the min and the max temperature with n bins. Play with the value of n and decides what value enables to best visualize the data. The normalization option 'probability' enable to display the percentage of data in each bin instead of the number of data points in each bin.

Do these histogram plots confirm the results of your linear regression?

2. Variations in trends

Increasing CO_2 concentration tends to cause an increase of global mean temperature because of the greenhouse effect. However, many other factors impact variations of global mean temperature. In particular, at decadal timescales, events such as El Niño or volcanic eruptions can cause fluctuations in temperature that will counteract or enhance the effect of rising CO_2. To quantify these fluctuations, calculate (again) trends in the temperature anomalies time series, but for each decade over the 1950-2016 period. For example, for the 1950's and for the global mean temperature anomaly time series, perform a linear regression of the 1950-1959 temperatures against time and find the slope, i.e., the temperature trend. Do this for all other decades and for the UBC temperature time series.

Hints: use masks to isolate temperatures and dates belonging to a given decade; you can use a for-loop to go through each decade. Check out the prelab if you don't remember how to use the regress() function and how to get slope estimate and 95% confidence intervals.

Report your results, as well as the trends over the entire period, in the first table of the Lab3 worksheet. Additionally, highlight the decades for which the trend is positive within the 95%

confidence interval (by circling them or underlining them) and do the same for the trend over the entire period. For example, if you estimate a value of 0.1 (you will have to report units in the worksheet) with a 95% confidence interval of [0.02 0.2], then the trend is positive within the 95% confidence interval. This would not be the case for an interval of [-0.02 0.2].

In general, are the local (UBC) or global trend the largest? The most significant? Did you find many decades for which the temperature trend is not positive within the 95% interval? What is the largest/smallest trend you found for each time series? Are you surprised by these numbers? Based on the trends you found and on your graph, would you say that there is an ongoing global warming trend since 1950? Is there any warming trend at UBC?

3. Local vs Global temperature

In the previous section, you may have found differences between trends in the temperature measured at UBC and global mean temperature. To further characterize these differences, plot the UBC temperature anomaly vs the global mean temperature anomaly (use markers symbol to plot, no continuous line). On your plot, annotate the correlation coefficient of the two time series as well as the associated p-value (you can use the function text() to do that.
Is there a significant correlation between both variables? Do you find the value of the correlation coefficient very large? Can you explain why the two time series are not perfectly correlated?

4. Impact of specific forcing on global temperature anomaly

For each of the 5 variables TSI, AOD, CO2, SO2 and MEI:
-perform a regression of the global mean temperature anomaly against the considered variable
-report the slope of the regression (with units!), the 95% confidence interval on the slope, and the coefficient of determination in table 2 of the worksheet

Do the sign of each slope match your expectations? Which variables seem to explain the most temperature variability? For which variable is there a large confidence on the sign of the linear regression slope?

In addition, make a figure where you plot the global mean temperature anomaly vs MEI and UBC temperature anomaly vs MEI on two different subplots. Perform a simple linear regression for each pair of variable and add the regression line and annotate the slope on each subplot. Do these scatter plot and the sign of the slopes match what you expected? Does the El Nino oscillation have a similar impact on global temperature and UBC temperature?

5. Combined impacts of multiple forcing on global temperature anomaly

You may have notice that none of the 5 explanatory variables (TSI, AOD, CO2, SO2 and MEI) used in this lab allow to explain most temperature variations. However, there combined effects may allow to explain a larger part of temperature variability. To assess this, perform a multilinear regression (cf prelab) where the dependent variable is the global mean temperature anomaly and the independent variables are the TSI, AOD, CO2, SO2 and MEI.
Report the coefficients and confidence interval associated to each independent variable in the prelab worksheet.

Additionally, plot the observed temperature against the temperature predicted by your multilinear regression. Add the line of equation y=x on your plot (all datapoints should be on this line if the regression was perfectly explaining variability in observed temperature) and annotate the coefficient of determination of the multilinear regression on your plot.

6. What can we conclude?

Before answering the wrap-up questions of the lab 3 worksheet, this section aims at highlighting some of the limitations of the methods used in this lab.

Load the data contained in 'dummyvariables_lab3.xlsx' available on connect.

First, plot the global mean temperature anomaly as a function of the age of Donald Trump. Calculate the correlation coefficient and pvalue of these two time series. Is there a strong correlation? What is the level of significance of the correlation (make them appear on your graph)? Is the increase in global mean temperature caused by Donald Trump growing older? Does this plot change your way of thinking about previous results from this lab?

Secondly, open a new figure that will be divided in two subplots.

- In the first one:
  - Plot the truncated CO_2 time series against the truncated temperature anomaly using marker symbols of your choice. Note that these data are **NOT real** and were made up from your TA imagination...
  - Perform a linear regression of temperature vs CO_2.
  - Annotate the slope, coefficient of determination and p-value on your graph
  - Plot the line corresponding to your regression using a continuous line. Do you think that your linear regression represent accurately the data?
- Then, on the second subplot:
  - Pot the temperature predicted by your linear regression for the complete CO_2 data (column five of 'dummyvariables_lab3.xlsx') using a continuous line.
  - On the same plot, plot the complete temperature anomaly against the complete CO2 data using marker symbols of your choice. About half of the data points are the **EXACT SAME** as the truncated data you plotted on the first subplot, but the complete data spans a larger change of CO2 values. Legend your graph. Again, note that these data are **NOT real**.

Does your linear model predict accurately the dependence of temperature on CO2? Why? How does it make you rethink our previous results? Can you think of a way to test the accuracy of the linear regression models used in previous part of the labs? What would be the limitation of such tests?