

Week 2: Lab exercises on multiple linear regression and stepwise regression

Exercise 1:

If you haven't done so already, run the scripts (MATLAB/Python) that were used for the examples in the lecture on Tuesday. See whether you can reproduce the figures from those examples (on synthetic and real data), and try to familiarize yourself with the codes.

Exercise 2:

Consider the following research problem:

Mountain glaciers worldwide are losing their mass in response to ongoing climate change. One of the main indicators of glacier 'health' is its mass balance. Annual glacier mass balance is a difference between total annual accumulation over the whole glacier surface (mass gain; mainly through snow accumulation) and total annual ablation over the whole glacier surface (mass loss; mainly through surface melting) over one year. Currently, almost all glaciers worldwide experience a negative mass balance, but their sensitivity to climate change (i.e. how much glacier mass is lost for a degree of warming) differs from glacier to glacier. In this project you are asked to use a data from a sample of mountain glaciers worldwide to investigate what drives the difference in their mass balance sensitivity to warming.

A good indicator of the mass balance sensitivity is a glacier mass balance gradient (Figure 1). Annual specific mass balance, b (expressed in meters water equivalent), derived for each elevation band along a glacier center-line, shows the following general pattern: lower elevations (glacier front or tongue) have negative b (more melting than accumulation over a year) while higher elevations of the same glacier have positive b (more accumulation than melting over a year).

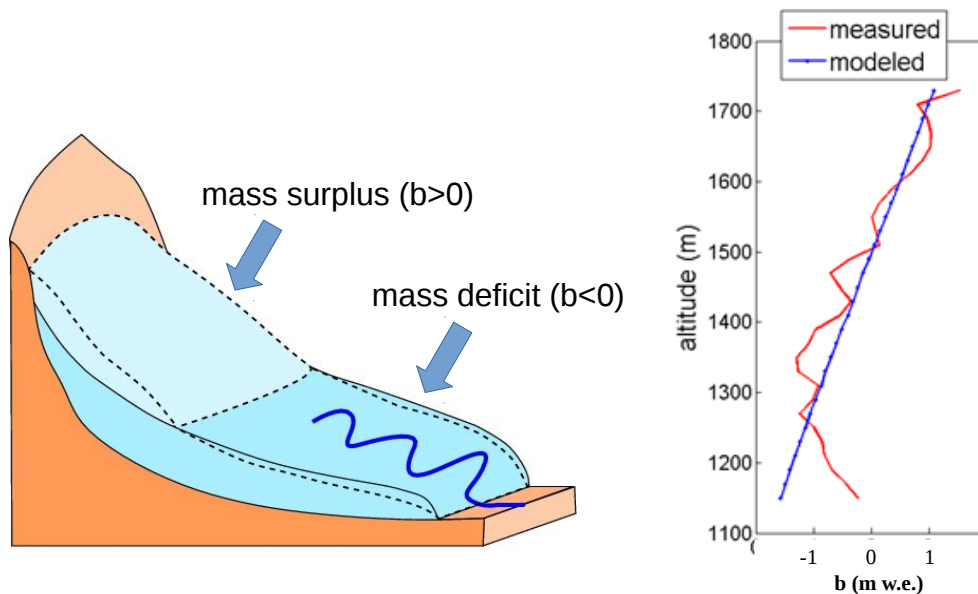


Figure 1: Left: A cartoon of a glacier showing the areas of mass surplus and mass deficit over a year. Right: An example of a measured annual mass balance profile: annual mass balance, b (m water equivalent), versus elevation, z (m above sea level). The slope of the fitted (modelled) line is the estimated mass balance gradient ($\Delta b / \Delta z$).

The mass balance profile with elevation is approximately linear (Figure 1), so a slope of the regression line between the mass balance and elevation gives a value of its mass balance gradient. Researchers found that, in general, the steeper the mass balance gradient (larger $\Delta b / \Delta z$) the more sensitive the glacier is to warming (local air temperature increase or decrease). But what dictates how steep the mass

balance gradient is? Local climate variables (e.g. annual temperature, precipitation) and topography (e.g. elevation of the terrain, slope of the glacier surface) are some of the prime suspects and your task today is to investigate which of these suspects can best capture the **spatial** variability in the glacier mass balance gradient.

You are given a dataset (file: **glaciers.mat** or **glaciers.csv**) collected for 136 glaciers worldwide. In mat file, the structure variable 'glaciers' consists of the estimated mass balance gradient ($g = \Delta b / \Delta z$) and other variables for each of 136 glaciers. Open **Lab2.m** or **Lab2_2020.ipynb** for more information on the data. We already wrote in some commands in these codes that plot the data, perform standardization of predictors etc. The codes with solutions are also provided (Lab2_solutions.m; Lab2_2020_SOLUTIONS.ipynb), but first try to work on the solutions yourself and use the solutions codes only if you get stuck.

Questions to be answered:

- 1) How well can the spatial variability in mass balance gradient be represented by a multiple linear regression (MLR) using all the plausible predictors from the given dataset? *[Hint: Squared correlation coefficient (r^2) between modelled and observed response variable tells how much variance is explained by the model.]*
- 2) What is the optimal number of predictors in MLR and how much variance in the mass balance gradient can be explained by this MLR? *[Hint: try both stepwise regression approach and the 'calibration-validation' approach similar to the one used in the Example 2 from Tuesday's lecture]*
- 3) Plot modelled (regressed) mass balance gradients versus observed ones on a scatter plot (with 1:1 line) and discuss whether the linear model represents a good fit to the data or not. *[Hint: in the scatter plot, check if there are any systematic biases from 1:1 line between modelled and observed b; are there any outliers]*