Atmospheric Modelling SS2019 - Exercise 3

“Statistical analysis tool: simple linear regression”

3.1 Introduction

This exercise is an introduction to linear modeling by the most basic model, the simple linear regression where an X variable is explained, modeled by an affine function of another variable Y. The purpose of such a model is multiple and therefore depends on the context and especially on the underlying issues. It may be just an exploratory approach or the search for an answer to a question of the type: does a quantitative variable X (e.g. *seasonal cycle or tendency or ENSO or all together*) have an influence on the quantitative variable Y (e.g. *mean age of air (AoA)*)? Or finally the search for a prediction model of Y as a function of X. Key concepts: model, estimates, tests, diagnoses are introduced and presented in this basic context. Their use and meaning depend on the objectives.

3.2 Regression model

Note Y the actual AoA variable to be explained and X the explanatory variable or fixed effect (*seasonal cycle, or tendency or ENSO or all together*). The model amounts to assuming that, on average, $E[Y]$, is an affine function of X. The writing of the model implicitly assumes a prior notion of causality in the sense that Y depends on X because the model is not symmetric.

$$Y = \beta_0 + \beta_1.X + \epsilon. \tag{1}$$

Or

$$E[Y] = f(X) = \beta_0 + \beta_1.X \tag{2}$$

$$\forall i = 1, ..., n \;\; E(\epsilon_i) = 0, \;\; Var(\epsilon_i) = \sigma^2 \tag{3}$$

3.3 Estimation of parameters

The estimation of the parameters $\beta_0, \beta_1, \sigma^2$ is obtained by maximizing the likelihood, under the assumption that the errors are Gaussian, or by minimizing the sum of the squares of the differences between observations and model (least squares). Both approaches lead to the same estimate while the maximum of the likelihood induces better properties of the estimators. For a sequence of observations $\{(x_i, y_i), i = 1, .., n\}$, the criterion of the least squares is written:

$$min_{\beta_0,\beta_1} \left(\sum_i^n (y_i - \beta_0 - \beta_1.x_i)^2 \right) \tag{4}$$

3.4 Applications in Python

The netcdf file contains timeseries of AoA, deseasonalized AoA, the seasonal cycle of AoA and the Multivariate ENSO Index.

3.4.1) Write a python script to read **AoA** variable contained into the netcdf file **(AoA_CLaMS_timeseries_era.nc)**.

3.4.2) Calculate the seasonal cycle of **AoA** and compare it with the variable **Seas_cyc** in the netcdf file.

3.4.3) Deseasonalize the **AoA** and compare it with the variable **AoA_deseas** in the netcdf file.

3.4.4) Detrend the deseasonalized **AoA** using simple linear fit (in python function **polyfit** of 1D).

3.4.5) From the deseasonalized and detrend AoA, estimate the impact of ENSO on the AoA. To do so, fit a line, $\mathbf{AoA\_deseas\_detr} = \beta\mathbf{.MEI} + \mathbf{c}$. We can rewrite the line equation as $\mathbf{AoA} = \mathbf{A.p}$, where $\mathbf{A} = [\mathbf{MEI\ 1}]$ and $\mathbf{p} = [\beta,\ \mathbf{c}]$. Now use the python function **linalg.lstsq** to solve for p.

3.4.5) Plot the **AoA**, deseasonalized and detrended **AoA**, the seasonal cycle, trend, ENOS-induced variation on AoA and the MEI.