Course title: EOTIST Standard course

Course subject: Modelling

Teacher: Marta Magnani

LESSON SM1.4
HOW TO BUILD A DATA-DRIVEN MODEL - EXERCISES

EOTIST project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 952111

[H2020 WIDESPREAD-05-2020 (Twinning)]

TABLE OF CONTENTS

Build quantitative models for terrestrial CO_2 fluxes 2

Exercise 1 2

Exercise 2 2

Exercise 3 3

Exercise 4 4

BUILD QUANTITATIVE MODELS FOR TERRESTRIAL CO_2 FLUXES

In this lesson we will build linear and non-linear models by using a freely available dataset that can be downloaded from Zenodo repository

direct link: https://zenodo.org/record/3588380#.Ymuru_NBzeq

The dataset contains measurements of fluxes (ER, GPP, NEE) and basic meteoclimatic variables (air temperature, air moisture, atmospheric pressure, soil temperature, soil moisture, solar irradiance) from 2017 to 2019. The fluxes were measured at the Nivolet Plain, Gran Paradiso National Park, western Italian Alps.

In the following we will explore Octave intrinsic functions (i.e. routines already implemented, native, in the software packages) that can be used to fit models to data in a broad range of applications.

EXERCISE 1

Familiarize with data:

1. Download the dataset and place it in a known working directory (e.g. 'Exercises_SM1_3/.').
2. Open a new script in Octave and save it in the same working directory of the dataset. Notice: points may not be accepted in script titles (e.g. 'Exercises_SM1.3.m' may produce a warning message).
3. Load the dataset in the script using the command

```
A=csvread('fluxes_meteoclimate_nivolet_V0.csv');
```

4. Select only fluxes and meteoclimatic variables (for simplicity use only the meteoclimatic variables flagged with '…_NEE') using

```
data=A(:,[7,9,11:4:33]);
```

5. compute GPP=NEE-ER and append the array to 'data'

```
gpp= A(:,7)- A(:,9);
data=[data,gpp];
```

   the last column of data contains the GPP now
6. Compute the Pearson correlation matrix using the command `corrcoef` (documentation at: https://octave.org/doc/v5.1.0/Correlation-and-Regression-Analysis.html)
7. What is the meteoclimatic variables that have larger correlation with the ER? And with GPP?
8. What meteoclimatic variables show larger correlation between them?

The correlation matrix is a symmetric table that contains the Person correlation coefficients. Such coefficients are used to assess collinearity of predictors: two predictors that have large correlation should not simultaneously included in a model.

EXERCISE 2

1. Try to build linear (univariate) models for both GPP and ER using the intrinsic function `regress` (documentation at: https://octave.sourceforge.io/statistics/function/regress.html) and the predictor

that showed larger correlations coefficients with respect to GPP and ER in Exercise 1 (excluding the other fluxes).

2. Build generalized linear models (GLM) using the same function and the 2 predictors that showed larger correlations coefficients with respect to GPP and ER in Exercise 1 (excluding the other fluxes).

EXERCISE 3

The air temperature and solar radiation are well known drivers of ER and GPP, respectively. The explicit functions read:

$$ER = a\, e^{bT}, \tag{1}$$

$$GPP = \frac{F\alpha\, R}{F+\alpha\, R}. \tag{2}$$

Here, the temperature *T* is expressed in Celsius, a is a free parameter, corresponding to the respiration at 0 °C, *b* is the temperature sensitivity of respiration, *R* is the incoming solar radiation, *F* is the maximum photosynthetic flux for infinite light supply and α is the apparent quantum yield, i.e. the photosynthetic response at low light level.

In GNU Octave framework, nonlinear models can be fitted with the intrinsic function `nlinfit` (documentation at: https://octave.sourceforge.io/optim/function/nlinfit.html). The fitting of data using a nonlinear model specified by the user is achieved by optimization techniques. It is therefore necessary to set starting values of the parameters to be fitted. These starting values are used as initial conditions for the optimization algorithms.

Try to fit the above nonlinear Equations using the following instructions:

1. Fit the nonlinear model in Eq. (1), setting

```
beta0=[1;1];
```

As starting values, and

```
model= @(b, x) (b(1) * exp (b(2) * x));
```

As model. Obtain the estimated parameters.

2. Try to fit the nonlinear model in Eq. (2), using

```
beta0=[-3;-0.003];
```

As starting values and

```
model=@(b, x) (b(1) * b(2) * x/( b(1) + b(2) * x));
```

As model. Obtain the estimated parameters.

Notice: in general, the convergence of the parameter values is not assured because the fit is not deterministic (i.e. in other contexts you may obtain extremely large, e.g. $O(10^{22})$, or negligible parameter values). Usually, good starting values facilitate the convergence. An indication of the parameter magnitude to be used as starting value is sometimes found in literature.

EXERCISE 4

Try to play with models:

1. Change the functional form. For instance, try to fit the equations

$$ER = (a_0 + a_1 VWC)\, e^{b_0 T_a}, \tag{3}$$

$$GPP = \frac{F_0 \alpha_0 R}{F_0 + \alpha_0 R}\,(A_0 + A_1 VWC)\ . \tag{4}$$

   Use for instance

```
model = @(b, x) (exp (b(1) * x(:,1))*(b(2)+b(3)*x(:,2)));
beta0=[1;1;1];
```

   for ER model and

```
model=@(b, x) ((b(1) * b(2) * x(:,1)/( b(1) + b(2) * …
x(:,1)))*(b(3)+b(4)*x(:,2)));

beta0=[-1;-0.0002;2;0.05];
```

   for GPP model.

2. Obtain the Akaike Information Criterion (AIC) using

$$AIC = N log\left(\frac{\sum_i \varepsilon_i^2}{N}\right) + 2k,$$

   Where ε_i are model residuals, *N* is the number of data and *k* the number of estimated parameters, including the residual variance (i.e. the number of parameters +1).

   Ref: Burnham,K.P.&Anderson,D.R.Apracticalinformation-theoreticapproach.*ModelSel.MultimodelInference***2**,70–71(2002).

3. Is the univariate regression in Exercise 3 better that the multi regression models fitted in this exercise? The best model should have lowest AIC.