Introduction to Econometrics

Summer 2020-1 In-Class Exercise, Lecture 1

We start with analyzing a dataset which is constructed "... to predict the burned area of forest fires, in the northeast region of Portugal, by using meteorological and other data". It is provided in Cortez and Morais (2007), and has been downloaded from `http://archive.ics.uci.edu/ml/datasets/Forest+Fires`.

(i) Find the documentation for the dataset and understand the meaning of each variable.
(ii) Do histograms of each variable, and get summary statistics of each of them. Can you detect any errors in the data?
(iii) Compute the correlation of each variable with the variable *area*.
(iv) Do a scatterplot of each variable versus the variable *area*.
(iv) Which variables do you think are main determinants of the burnt area? After you decide on a few, check whether they are highly correlated with the other variables.

References

Cortez, P. and A Morais (2007), "A Data Mining Approach to Predict Forest Fires using Meteorological Data," N Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence.