T3. Models

Models

Blas MOLA-YUDEGO

Summary

What is a model? How is a linear regression fit and assessed? How we construct a model and make predictions with it?

In our case, a model is simply a (simplified) representation of reality, using mathematical language. This section will deal with regression linear models, starting with a single variable (predictor, independent) used to predict another one (response, dependent). The relation between both variables must be modeled, using a line and some properties of the normal distribution to fit the parameters that define that line. The model is assessed to measure the predictive power as well as if we incur in any violation of the premises concerning the way the line was fit. We will review the assumptions of Normality of the residues, linearity, constant variance (*homoscedasticity*) and independence. From that basis, we will expand to add more variables, check the effects of *multi-collinearity* and how to deal from there.

Materials

T3.1 Linear regression model [lecture]

T3.2 Model assessment [lecture]

T3.3 Multiple regression [lecture]

Slides from old lectures 2019 [PDF]

Tasks

We propose you to try the following tasks to practice the concepts explain in those lectures:

1. Create a large sequence of numbers following a normal distribution with defined **mean** =0 and **st d deviation ()** using excel/R. That will be the noise in the model.
2. Create a sequence of numbers, either random, systematic (e.g. 1 to 100) or following a normal distribution. That will be the **x** in the model.
3. Create a model. For instance y=2+3x. In this case, 0=2 and 1=3. This is the true model of your data. If you try to make a figure, it will look like a perfect line, with that exact formula and R2=1
4. Add the noise. That is, to add to y=2+3x the values of step 1.
5. Now check how the model behaves in the figure. Increase the noise (increase the **std deviation ()** of step 1). How is the R2 changing? Are you being fooled by randomness? Do you see a "better picture" with a larger sample?

How to do it?

In Excel:

Generate random numbers in excel: =RAND()
Generate numbers following a normal distribution with *mean*=**100** and *st dev*=*10*: =NORMINV(RAND(), 100,*10*)

In R:

Check here.

For more instructions, **google** (as I do)!

Objectives

The objectives of this topic are:

- To understand the main assumptions of linear regression models
- To fit linear regression models
- Model assessment and evaluation

Schedule

Lecture days: Third week: 2-6 November

During the lecture we will discuss about the topics in the presentations, the exercises and possible doubts.

Data and materials

Datasets

Excel for task [xls]

Excel for practice [xls]

Simple regression exercises

Exercise pre-exam [PDF]

Exercise height-volume [PDF]

Exercise heigh-diameter-volume [PDF] [solutions]

Exercise wheat production [PDF]

Multiple regression exercises

Exercise height-diameter-value [PDF]

Exercise barley yields [PDF]

Videos and tutorials

Simple linear regression [youtube]

How to make our own *sandbox* model [video]

Reflections

What are the consequences of violating the main assumptions of linear regression models?

What is the role of the interception ($_0$) in a model?

How do you decide the variables to be included in a model?

[These questions may help the students to focus and reflect on the topic contents. They do not require to be submitted as an assignment and are not to be evaluated.]

"All models are wrong. Many are useful. Some are lethal" (Taleb, echoing George Box)