Some important concepts: factor, covariate, ANOVA, regression and ANCOVA

Souvik Roy

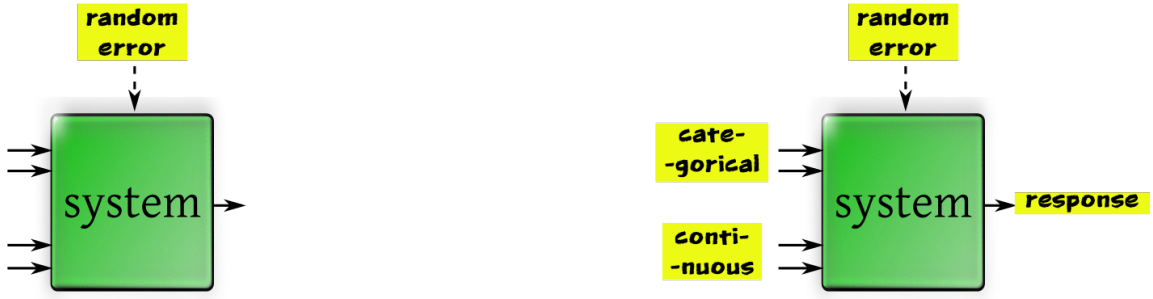
1 Introduction

Suppose we have a machine that takes several inputs and returns possibly several outputs. Since no machine is perfect, in the sense that it is not absolutely efficient, it is expected that for some inputs, the machine might produce outputs that are different from the desired ones. It is reasonable to consider this behaviour random, under the assumption that the machine is manufactured ideally, and is not biased against inputs. Thus the error produced by the machine is random. Now, we shall simplify our machine and add in the assumption that our machine takes in inputs that are either numbers (and hence we can do all sorts of arithmetic operations with them[1]) or quantities that cannot be compared in the given context (like name and gender [2]). The numerical inputs are usually assumed to be real numbers and since the real number line is continuous, they are known as **continuous variables**. The *non-comparable* inputs are known as **categorical variables** since they usually represent categories in which the data is or can be grouped in some sense Thus, for our new highly simplified machine (which we call a **system** here), we have a structure that is succinctly represented by figure 1 on the next page.

Now we perform some renaming. The categorical variables will be henceforth called **factors** and the continuous inputs will the called **covariates**. The reason for this renaming might not be clear at the moment but, we shall soon see why these names are a much better choice. In the present scope we shall consider the output (or the *response* as in the figures) to be continuous as well. If we allow the output to be categorical, or maybe a combination of both categorical and continuous, the **system** here will be called a **generalized linear model**. If however we have a continuous output,

[1]Of course we mean operations that make sense given the context.

[2]Context is necessary here as well, for instance, monthly income might not be comparable when we discuss about genetic diseases, but that is not the case as soon as we consider monthly expenditure.



(a) Our machine

(b) Our machine with the inputs

Figure 1

the model is just a **linear model**. It is to be noted that the random error is in both the cases, always taken to be continuous.

2 Several terminologies

Here we introduce some additional terms that will be required in this course.

- Suppose in our machine, there are only factors and no covariates in the input, for example a system where we work with two drugs and gender, and the machine being the patient. Clearly, there are no covariates in this setup. Such a linear model is called an **ANOVA model**, which is the acronym for Analysis of Variance. It shall later be clear why it is named this. If there are k factors, then the model will be called a k - *way ANOVA*. For example, a linear model with only one categorical input will be called a one-way ANOVA.
- Now consider a linear model in which there are only covariates in the input. Such a model is called **regression**. The term regression can have implications far beyond the scope of linear models (hence it might be more appropriate to say linear regression instead), but in general it is used to refer to a model that takes only covariates as inputs.
- Lastly, if we have a combination of factors and covariates in our input, then, the corresponding linear model is called **ANCOVA** which is the acronym to Analysis of Covariance. This is the most general model of all the three mentioned here, and in this course we shall essentially discuss this model. Here too, we shall have ways, that is, an ANCOVA model is called a *k-way*

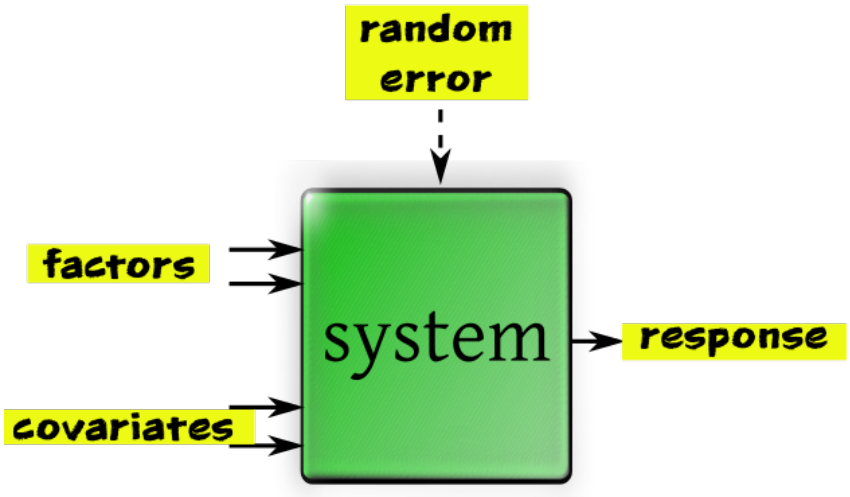


Figure 2: Factors and Covariates

ANCOVA model if the **number of factors is *k***. It is important to note that we consider the number of factors and **not the number of covariates** while naming this model as a *k-way ANCOVA*.

It is important to note that in each of the above cases, the random error and the response are always taken to be continuous. After renaming, the system looks like figure 2