# Universistat Politècnica de Calatunya

## Statistical Modelling and Design of Experiments

Master in Innovation and Research in Informatics

# Assignment 2

*Author:*
First name Daniel
Pujazon Bonet

June 14, 2021

# 1 Generate your data.

## 1.1 Define, for each factor (from 1 to 5) a distribution (the RVGs that you prefer, uniform, normal, exponential, etc.). For the factors 6 to 10 define a function that uses the previous variables, as an example F6=F1+2F3.

## 1.2 Define an answer variable that will be composed by a function that combines a subset of the previous factors plus a normal distribution you know (to add some random noise).

The data stream is generated through the attached DataGen.cpp file and it's save it on a .csv file called DataSet.csv (also there's a .txt file).
Take note that factors indexing (nomenclature) will start from 0. So the first 5 factors, defined by probability distributions, are F0, F1, F2, F3 and F4. The rest F5, F6, F7, F8, F9 and F10 are the factors resulting on a linear combination of the first 5 factors being the F10 the ASNWER output variable.

# 2 Obtain an expression to generate new data.

**Imagine that you don't know nothing regarding how this dataset has been generated. Consider that the factors represent different machines and the answer is the time to do an operation.**
**You need to explore it because you want to define a model to obtain new data for your DOE (you want to detect the possible relations and the interactions between the factors, or maybe you want to test alternatives or predict future scenarios).**

## 2.1 Explore the possible relations of all the factors and the answer variable, you can use any technique developed during the course (LRM or ANOVA).

We are gonna to do a PCA due to know the iterrelation between the different factors. After that we will generate some Linear Regresion models (LRM) and we will take the one that fits better our requirements. All will be made through the R script ObtainExpresion.r
So, applying PCA to the DataSet.csv we have found:

- F3, F8, F6, F1, F7 and F2 are strong correlated. F3,F8 and F6 are positive correlated while F1, F7 and F2 are negative correlated.
  On the other side, F0, F5, F4 and F9 are strong and positive correlated.

- Each group angle respect to one of the axis is near to 0, what means that these group is fully explained only by one PC: First group (F3,F8...) is almost parallel to y-axis, PC2, while second one (F0, F5...) is almost parallel to x-axis, PC1. This means that the other-PC group variables has no effect (because they're independant or because the impact of the other variable on the result it's too weak) on the value of the ohter-PC group vairables (pe: F5 value is not significantly determined by F3, F6 is not determined by F0... and so on)

- Scree plot show us that we need at least 4 variables to explain more than the 80% of the system variables variation (even if we go further, with 5 variables we explain 100% of variability so this means that from 6th variable, which explain 0% of the variance, are full explained by the other variables).

So, knowing that there are 5 variables that explain 100%, we could do all the possible LRM using 5 of the 10 variables (which would be 252, combination without repetitions of 10 elements taken from 5 to 5) and we will get the one that fit better.

A LRM with all variables of course exlpain 100% of the variation on answer. A LRM2 with all the F1,F2,F3,F4,F5 (without F0) variables explain also the 100%. On LRM3 if we only take the first 4 factors (as scree plot conclusions). With that we only can explain the 78%, so we will use the or LRM2 or LRM4. Take a look that p-value is higher than accepted value so

we won't use it:
$(LRM2)Answer = +5(X1) - 3(X2) - 5(X3) - 1.18 * 10^{-6}(X4) + (X5)$
$(LRM4)Answer = +4(X0) + 5(X1) - 3(X2) - 5(X3) + (X4)$

If we compare both, all variables on both LRM have the same p-value (so more or we see that LRM2 X4 has bigger p-value than X3 on LRM3. Also residual standrad error on LRM4 is smaller than LRM2, so we would pick LRM2 instead of LRM4. The last check between two models can be apply the inputs from DataSet and compare both outputs with the DataSet answer. This is made on DataGen.cpp and dumped on testLRM.txt. We can see how LRM4 outputs are more near to answer than LRM2. So let's use LRM4.

### 2.2 Describe what you find on this analysis and, explain if it is coherent with the knowledge you have from the data.

(Confirm it, both PCA and LRM doing "test", input the data of 5 factors and get the output).

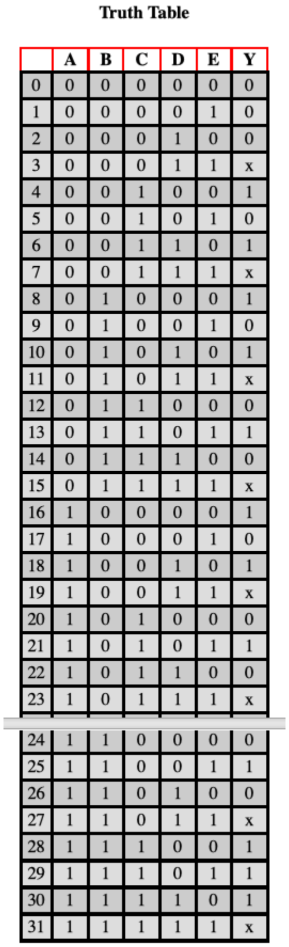
- First of all, the scree plot is coherent with the fact that from Factor6 to Factor10 are totally and only dependant on the first 5 factors.

- Also is coherent that F8,F6 and F7 are strongly correlated with F3, F2 and F1. Their definitions are: F6=$F1 + 3 * F3$; F7=$2 * F2 + 3 * F2$; F8=$F3 - F1$
  The signs inside each function also explains why, for example, F8 is negative correlated with F1.

- On the same way correlation between F5, F9 with F0 and F4 make sense: F5=$4 * F0 + F4$; F9=$F4 - F0$.

- If we calculus theoretically the Answer function, this is X which can be decomposed on the independant factors, we have the same one as the LRM4.

### 2.3 Use a simulation model to generate new data. The simulation model will be a very simple model composed by one server by each one of the factors you use on the answer.



**Truth Table**

| | A | B | C | D | E | Y |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 1 | x |
| 4 | 0 | 0 | 1 | 0 | 0 | 1 |
| 5 | 0 | 0 | 1 | 0 | 1 | 0 |
| 6 | 0 | 0 | 1 | 1 | 0 | 1 |
| 7 | 0 | 0 | 1 | 1 | 1 | x |
| 8 | 0 | 1 | 0 | 0 | 0 | 1 |
| 9 | 0 | 1 | 0 | 0 | 1 | 0 |
| 10 | 0 | 1 | 0 | 1 | 0 | 1 |
| 11 | 0 | 1 | 0 | 1 | 1 | x |
| 12 | 0 | 1 | 1 | 0 | 0 | 0 |
| 13 | 0 | 1 | 1 | 0 | 1 | 1 |
| 14 | 0 | 1 | 1 | 1 | 0 | 0 |
| 15 | 0 | 1 | 1 | 1 | 1 | x |
| 16 | 1 | 0 | 0 | 0 | 0 | 1 |
| 17 | 1 | 0 | 0 | 0 | 1 | 0 |
| 18 | 1 | 0 | 0 | 1 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 | 1 | x |
| 20 | 1 | 0 | 1 | 0 | 0 | 0 |
| 21 | 1 | 0 | 1 | 0 | 1 | 1 |
| 22 | 1 | 0 | 1 | 1 | 0 | 0 |
| 23 | 1 | 0 | 1 | 1 | 1 | x |
| 24 | 1 | 1 | 0 | 0 | 0 | 0 |
| 25 | 1 | 1 | 0 | 0 | 1 | 1 |
| 26 | 1 | 1 | 0 | 1 | 0 | 0 |
| 27 | 1 | 1 | 0 | 1 | 1 | x |
| 28 | 1 | 1 | 1 | 0 | 0 | 1 |
| 29 | 1 | 1 | 1 | 0 | 1 | 1 |
| 30 | 1 | 1 | 1 | 1 | 0 | 1 |
| 31 | 1 | 1 | 1 | 1 | 1 | x |

Then, the first thing we have to do is define the maximum and minimum values for each factor. We will take the maximum and minimim values, (F5 has significant outliers so we take the minimum and maximum value at least with a 1% of frequency to appear), from the empricial results (even we could also taken knowing the potential values that each distribution can generate. This is on ObtainExpresion.R):

- For Factor 1: (+):= 89.73; (-):= 12.50.
- For Factor 2: (+):= 00.95; (-):= 6.25e-07.
- For Factor 3: (+):= 49.97; (-):= 20.02.
- For Factor 4: (+):= 02.16; (-):= 1.37.
- For Factor 5: (+):= 31.1965; (-):= -20.162.

This means that, at least with our current design, we have 32 scenarios (on the truth table the '0' means the minimum value, (-), while the one is the maximum, (+) and A,B,C,D,E are the F0, F1, F2, F3, F4 factors respectively).
Knowing that we implement our model using GPSS the simulation model. It's the attached file M2.gps and has been deployed with GPSS World.

The idea is, you define which scenario are you going to simulate on the

SCENARIO variable according to the truth table (pe: SCENARIO 3 means F0 (-), F1 (-), F2 (-), F3 (+), F4(+). Then there's a function for each Factor which has the value, already calculated, for both (+) and (-) scenarios. First attempt was using queues (interpreting factors values as time) but distributions give us negative times so it make no sense. Then it's assigned to a ANSWER variable and dumped. Also added uniform random noise.

# 3 DOE

**Now you have a model to generate new data. This model can be used to generate data for the different scenarios that must be considered.**

## 3.1 Define a DOE to explore with what parametrization of the 10 factors the answer obtains the best value (define what means best, i.e. maximize or minimize the value).

## 3.2 Detect and analyze the interactions.