Design of Experiments: introduction

The study of **Design of Experiments** (DOE) involves the systematic planning, execution, and analysis of experiments to understand and optimize the relationship between input variables (factors) and the output response. It aims to identify the most significant factors affecting the response, improve process efficiency, and make data-driven decisions to achieve desired outcomes in various fields, i.e science, engineering, and industrial applications.

A **Factorial Design** is an experimental design used in statistics and experimental research to study the effects of two or more independent variables, also known as factors, on a dependent variable.

A **Central Composite Design** (CCD) is a type of experimental design commonly used in response surface methodology (RSM). It is a special type of factorial design that allows for the exploration and optimization of response surfaces in a systematic and efficient manner.

Example of 2^2 factorial design

(After Montgomery D., Design and Analysis of-Experiments, 2012) We consider the effect of the concentration of the reactant (factor A) and the amount of the catalyst (factor B) on the conversion (yield) in a chemi- cal process. The two levels of factor A are 15 and 25 percent. The two levels of factor B are 1 and 2 (pounds). The exper iment is replicated three times for each combination. The data obtained are as follows:

Factor		Combination	Replicate			Total
-	-	A low, B low	28	25	27	80
+	-	A high, B low	36	32	32	100
-	+	A low, B high	18	19	23	60
+	+	A high, B high	31	30	29	90

Estimation of the average effects

$$A = \frac{1}{2(3)}(100 + 90 - 80 - 60) = 8.333333$$

$$B = \frac{1}{2(3)}(-60 + 90 - 80 - 100) = -5$$

$$AB = \frac{1}{2(3)}(80 + 90 - 100 - 60) = 1.666667$$

The effect of A (reactant concentration) is positive; this suggests that increasing A from the low level (15%) to the high level (25%) will increase the yield. The effect of B (catalyst) is negative; this suggests that increasing the amount of catalyst added to the process will decrease the yield. The interaction effect appears to be small relative to the two main effects

Fitting a linear model with an interaction

A linear model without an interaction would be of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

A linear model with an interaction would be of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

The output for example for the second model is given below. We see that the interaction is not significant and that the standard error is minimal and constant for all estimates. In the next two slides we give the R and Python code to compute such models.

```
# Residual Standard Error=1.9791
# R-Square=0.903
# F-statistic (df=3, 8)=24.8227
# p-value=2e-04
#                  Estimate Std.Err t-value Pr(>|t|)
# Intercept         27.5000  0.5713 48.1354   0.0000
# Reactant           4.1667  0.5713  7.2932   0.0001
# Catalyst          -2.5000  0.5713 -4.3759   0.0024
# ReactantCatalyst   0.8333  0.5713  1.4586   0.1828
```

R code

```r
# Create a data frame
data <- data.frame(
  Factor = c("- -", "+ -", "- +", "+ +"),
  Combination=c("A low, B low","A high, B low","A low, B high","A high, B high")
      ,
  Reactant = c(15, 25, 15, 25),
Catalyst = c(1, 1, 5, 2),
  Replicate_A = c(28, 36, 18, 31),
  Replicate_B = c(25, 32, 19, 30),
Replicate_C = c(27, 32, 23, 29),
  Total = c(80, 100, 60, 90),
AverageYield = c(26.7, 33.4, 20,30))

datacodedminusplus <- data.frame(
  react = c(-1,+1,-1,+1,-1,+1,-1,+1,-1,+1,-1,+1),
  cata = c(-1,-1,+1,+1,-1,-1,+1,+1,-1,-1,+1,+1),
  yield = c(28,36,18,31,25,32,19,30,27,32,23,29))

Reactant <- datacodedminusplus[,1]
Catalyst <- datacodedminusplus[,2]
ReactantCatalyst <- datacodedminusplus[,1]*datacodedminusplus[,2]

datacoded <- data.frame(Reactant, Catalyst, ReactantCatalyst)
ls.print(lsfit(datacoded, datacodedminusplus[,3])) # only the intercept is
      significant
```

Python code

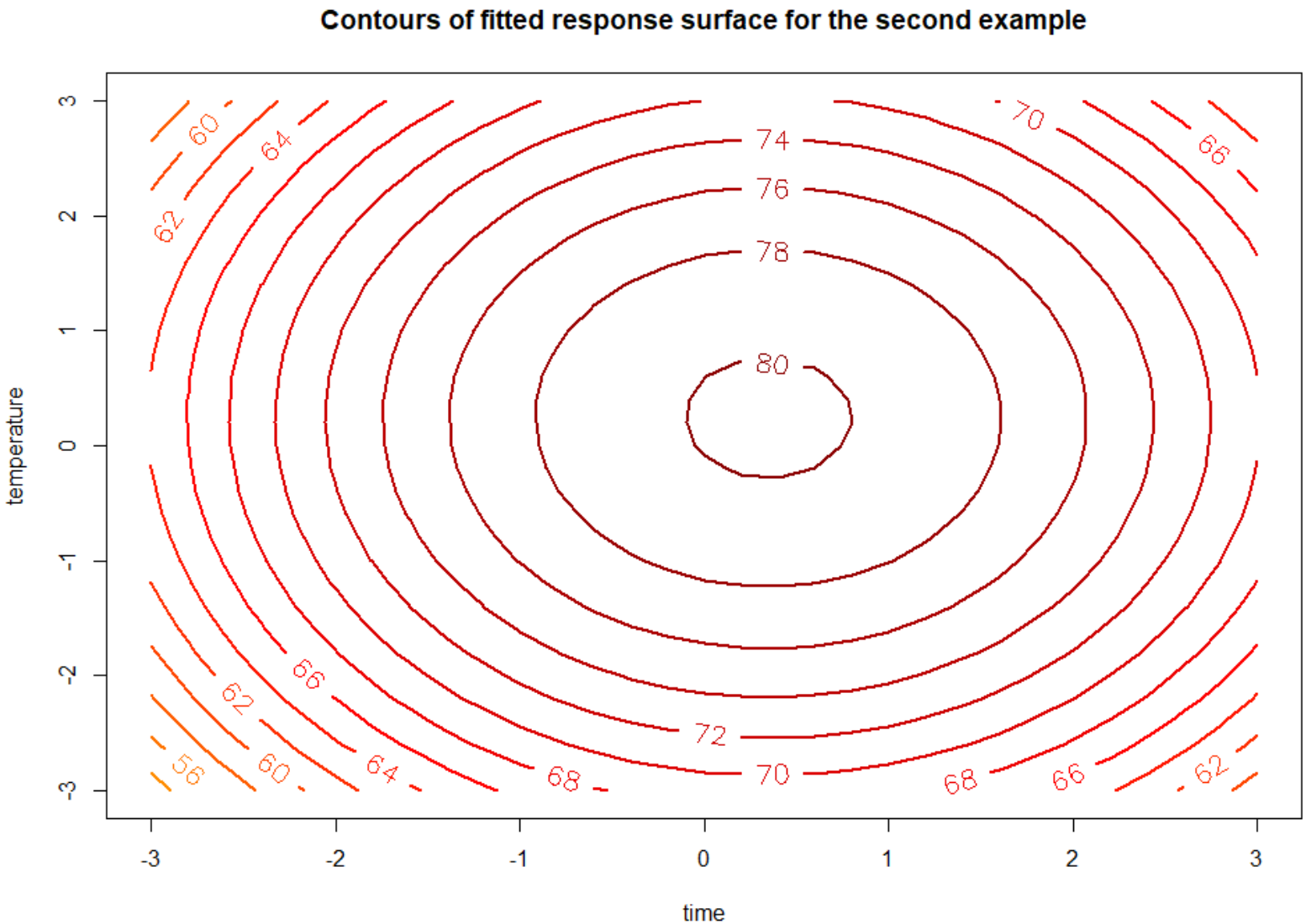
```
1  import pandas as pd
2  import statsmodels.api as sm
3
4  # Create a data frame
5  data = pd.DataFrame({
6      'Factor': ['- -', '+ -', '- +', '+ +'],
7      'Combination': ['A low, B low', 'A high, B low', 'A low, B high', 'A high, B
         high'],
8      'Reactant': [15, 25, 15, 25],
9      'Catalyst': [1, 1, 5, 2],
10     'Replicate_A': [28, 36, 18, 31],
11     'Replicate_B': [25, 32, 19, 30],
12     'Replicate_C': [27, 32, 23, 29],
13     'Total': [80, 100, 60, 90],
14     'AverageYield': [26.7, 33.4, 20, 30]})
15
16 datacodedminusplus = pd.DataFrame({
17     'react': [-1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1],
18     'cata': [-1, -1, 1, 1, -1, -1, 1, 1, -1, -1, 1, 1],
19     'yield': [28, 36, 18, 31, 25, 32, 19, 30, 27, 32, 23, 29]})
20
21 Reactant = datacodedminusplus['react']
22 Catalyst = datacodedminusplus['cata']
23 ReactantCatalyst = datacodedminusplus['react'] * datacodedminusplus['cata']
24
25 datacoded = pd.DataFrame({'Reactant': Reactant, 'Catalyst': Catalyst, '
       ReactantCatalyst': ReactantCatalyst})
26 X = sm.add_constant(datacoded[['Reactant', 'Catalyst', 'ReactantCatalyst']])
27 y = datacodedminusplus['yield']
28
29 model = sm.OLS(y, X).fit()
30 model.summary()
```

Another example

We consider the yield of a chemical process optimized in function of time and temperature. This is an example of a factorial design with three replications at the center (runs 1 through 7), which has been augmented to a Central Composite Design (CCD) with four additional star points (runs 8 through 11).

Run	Time	Temperature	X1	X2	Yield
1	80	170	-1	-1	76.5
2	80	180	-1	1	77
3	90	170	1	-1	78
4	90	180	1	1	79.5
5	85	175	0	0	79.9
6	85	175	0	0	80
7	85	175	0	0	80.3
8	92.07	175	1.414	0	8.45
9	77.93	175	-1.414	0	75.6
10	85	182.07	0	1.414	78.5
11	85	167.93	0	-1.414	77

Contours of the fitted response surface



Conclusions

The full second order linear model computed on the Central Composite Design and for which we have drawn the contours of the response surface is the following:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

We see that with a minimum number of experiments and therefore cost for development (only 11 runs), we are able to have a precise idea of the optimal combination of factors (time and temperature) that gives the maximum yield of the chemical process. Time is therefore optimal between 85 and 90 and tem temperature should be 175.

Traditional optimization methods are helpful in the first place to know which values of the factor lie close to the optimal regions, for example the simplex algorithm or the steepest ascent method.

References

Montgomery D., Design and Analysis of Experiments, 2012, Wiley

The R Project for Statistical Computing:
https://www.r-project.org/

Python:
https://www.python.org/

course notes