Modelling and design of experiments

Chapitre 3: Analysis of variance

Dr Jean-Marie Fürbringer

École Polytechnique Fédérale de Lausanne

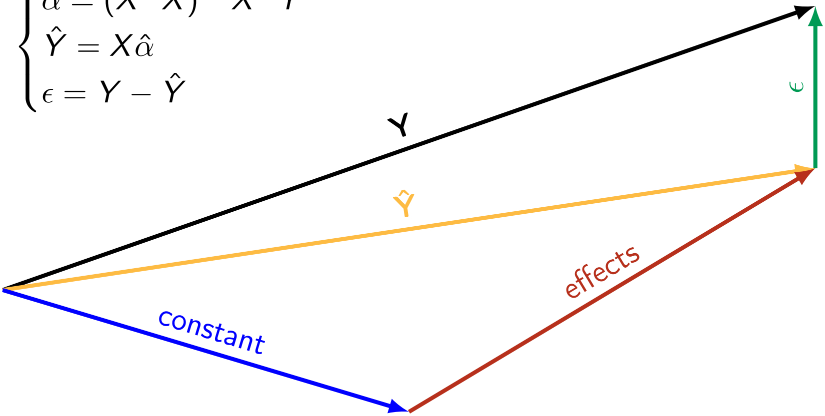
Fall 2024

3.1.1 Pedagogical objectives

1. Understand how an ANOVA table is structured
2. Understand the consequences of non orthogonality
3. Being able to compute an ANOVA table for orthogonal and non-orthogonal designs
4. Being able to use and interprate the Matlab routine *fitlm()*
5. Being able to interprate the output of a regression

3.1.2 Analysis of variance (geometric perspective)

$$\begin{cases} \hat{\alpha} = (X^T X)^{-1} X^T Y \\ \hat{Y} = X\hat{\alpha} \\ \epsilon = Y - \hat{Y} \end{cases}$$



3.1.3 ANOVA of a model as a whole

Sources	SS	DF	MS	F	p
Model	$SS_{\hat{Y}}$	P	$MS_{\hat{Y}} = \frac{SS_{\hat{Y}}}{P}$	$x = \frac{MS_{\hat{Y}}}{MS_{\epsilon}}$	$F(x, P, N-P)$
Residue	SS_{ϵ}	$N-P$	$MS_{\epsilon} = \frac{SS_{\epsilon}}{N-P}$		
Total	SS_{Y}	N	–		

N is the number of runs and P, the number of coefficients in the model

3.1.4 ANOVA of the Young modulus model

Sources	SS	DF	MS	F	p
Model	396 990.2	4	99 247.55	$4.8\ 10^{6}$	$1.22\ 10^{-16}$
Residue	0.1	5	0.02		
Total	396990.3	9	–		

Model : $E = 210 + 0.24x_C - 0.63x_S - 0.053x_T$ avec $x_i \in [-1, 1]$
99% of the SS comes from the constant : So this table does not give interesting information on the quality of the model.

3.1.5 ANOVA without the constant

Sources	SS	DF	MS	F	p
Model (without const.)	2.00	3	0.67	32.2	0.11%
Residue	0.1	5	0.02		
Total	2.10	9	–		

The analysis without the constant is *sharper*, indicating clearly that the experiments have put effects in evidence

It would be also interesting to know which coefficients of the model are significant and which one could be neglected (parsimony principle).

3.2.1 How to decompose a model

At the level of the linear system, the parting of a model in two sub-models is done that way :

$$\begin{aligned} \hat{Y} &= \hat{Y}_1 + \hat{Y}_2 \\ X\hat{\alpha} &= [X_1\ X_2]\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = X_1\hat{\alpha}_1 + X_2\hat{\alpha}_2 \end{aligned}$$

At the level of the sum of squares it gives :

$$\begin{aligned} \hat{Y}^2 &= (\hat{Y}_1 + \hat{Y}_2)^2 \\ &= \hat{Y}_1^2 + 2\ \hat{Y}_1 \cdot \hat{Y}_2 + \hat{Y}_2^2 \\ &= \hat{Y}_1^2 + \hat{Y}_2^2 \quad \text{if and only if } \hat{Y}_1 \cdot \hat{Y}_2 = 0 \end{aligned}$$

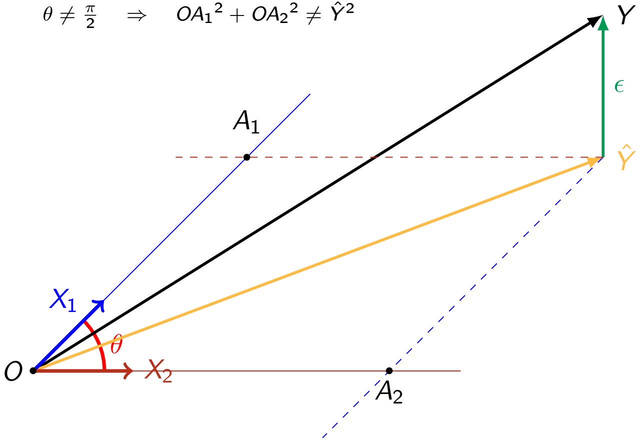
3.2.2 ANOVA for two orthogonal parts

Source	SS	DF	MS	F	p
Partie 1	$SS_{\hat{Y}_1}$	P_1	$\frac{SS_{\hat{Y}_1}}{P_1}$	$x_1 = \frac{MS_{\hat{Y}_1}}{MS_\epsilon}$	$F(x_1, P_1, N-P)$
Partie 2	$SS_{\hat{Y}_2}$	P_2	$\frac{SS_{\hat{Y}_2}}{P_2}$	$x_2 = \frac{MS_{\hat{Y}_2}}{MS_\epsilon}$	$F(x_2, P_2, N-P)$
Résidu	SS_ϵ	$N-P$	$\frac{SS_\epsilon}{N-P}$		
Total	SS_Y	N	–		

N is the number of runs and P_1 et P_2, the number of coefficients of the parts 1 and 2 respectively, $P = P_1 + P_2$

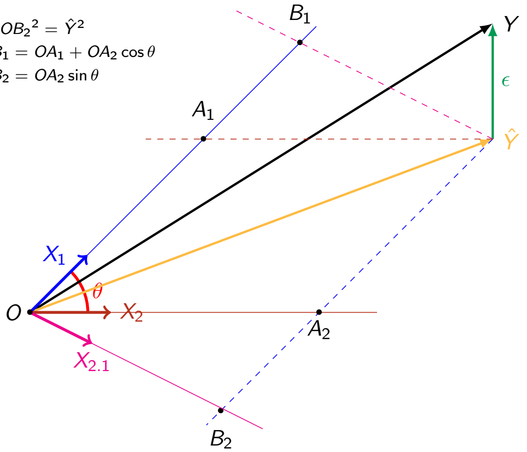
3.2.3 Orthogonal decomposition

$$\theta \neq \frac{\pi}{2} \quad \Rightarrow \quad {OA_1}^2 + {OA_2}^2 \neq \hat{Y}^2$$



3.2.3 Orthogonal decomposition (2)

$$\begin{cases} {OB_1}^2 + {OB_2}^2 = \hat{Y}^2 \\ a_1^* = OB_1 = OA_1 + OA_2 \cos\theta \\ a_2^* = OB_2 = OA_2 \sin\theta \end{cases}$$



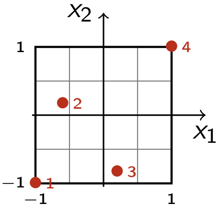
3.2.4 Determination of a cocktail effect

Toxicity of two combined products

- **Objective**: Identify the interaction of the toxicity of two products P_1 and P_2 found in 4 sodas
- **Type of investigation**: Kind of observation (factors can not be freely manipulated
- **Factors**:
  1. P_1 concentration, $C_1 \in [0.1g/l, 0.6g/l]$
  2. P_2 concentration, $C_2 \in [0.08g/l, 0.2\%]$
- **Response(s)**: Measurement of an enzyme production related to the risk of diabet
- **Model(s)**: Empirical linear model with interactions $Y = a_o + \sum_i a_i x_1 + \sum_{i<j} a_{ij} x_i x_j$
- **Strategies**: Perform tests with commercialized sodas

3.2.5 Design of experiments

- Points of measurement :



- Model matrix :

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -0.6 & 0.17 & -0.1 \\ 1 & 0.2 & -0.83 & -0.17 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -0.6 & 0.17 & -0.1 \\ 1 & 0.2 & -0.83 & -0.17 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

- Dispersion matrix :

$$(X'X)^{-1} = \begin{pmatrix} 0.22 & -0.01 & 0.07 & -0.19 \\ -0.01 & 0.39 & -0.25 & 0.02 \\ 0.07 & -0.25 & 0.36 & -0.08 \\ -0.19 & 0.02 & -0.08 & 0.41 \end{pmatrix}$$

- Variance inflation factors :

-	a_0	a_1	a_2	a_{12}
VIF	1.7	1.9	2	1.7

The analysis shows that the design is applicable

3.2.6 Inference of the coefficients

After the experiments

Experimental data :

Expériences	1	2	3	4
Y(set 1)	80.4	70.8	67.1	270.0
Y(set 2)	89.7	58.9	53.7	275.3

Model coefficients :

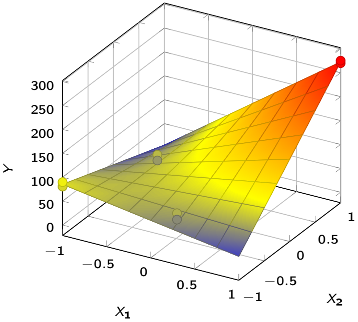
-	a_0	a_1	a_2	a_{12}
α_i	97.7	52.6	41.2	81.2
α_i/α_o	-	54%	42%	83%

Estimator :

$\hat{\alpha} = (X'X)^{-1}\ X'Y$

Model :

$Y = 97.7 + 52.6x_1 + 41.2x_2 + 81.2x_1x_2$



3.2.7 Angles between regressors and SS

- The information matrix $(X'X)$ gives the product of the regressors 2 by 2
- The scalar product is defined ase $\vec{x}_i \cdot \vec{x}_j = \|\vec{x}_i\| \|\vec{x}_i\| \cos\phi_{ij}$
- The angles between the regressors can then be computed by $\phi_{ij} = \arccos\left(\frac{k_{ij}}{\sqrt{k_{ii}}\sqrt{k_{jj}}}\right)$ if k_{ij} are the element of the matrix of information

-	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_{12}}$
I	97^o	102^o	53^o
$\mathbf{x_1}$	-	47^o	89^o
$\mathbf{x_2}$	-	-	87^o

Regressor	I	x_1	x_2	x_{12}
$SS(a_i x_i)$	76 366	13 291	9 231	26 842
$\sum SS(a_i x_i)$		125 730		
SS(Y)		178 863	$R = 1.42$	

3.2.8 Sequential orthogonalisation

1. Compute the half effect, the estimate and the residue for a model with 1 regressor (let's say a_o)
2. Compute the sum of squares $SS(a_o)$ for this model and $SS(\epsilon_o)$ for the corresponding residue
3. Compute the half effects, the estimates and the residue for a model with 2 regressors (let's say a_o et a_1)
4. Compute the sum of squares $SS(a_1|a_o)$ by subtracting $SS(a_o)$ from the sum of squares of the model with two regressors (or from the difference between the sums of squares of the two residues
5. etc.

3.2.9 ANOVA with SS of type I

- $SS(\hat{Y})$
- $SS(a_o)$
- $SS(a_1|a_o)$
- $SS(a_2|a_o, a_1)$
- $SS(a_{12}|a_o, a_1, a_2)$

Source	SS	SS*	DF	MS	F	P
Model	178 863		4	44 716	818	$1.1\ 10^{-8}$
Residue 1	218		4	55		
a_o		116 634	1			
Residue 2		62 447	7			
a_o		116 634	1			
a_1		37 034	1			
Residue 3		25 414	6			
a_o		116 634	1			
a_1		37 034	1			
a_2		9 139	1			
Residue 4		16 274	5			
a_o		116 634	1	116 634	2135	$1.23\ 10^{-9}$
a_1		37 034	1	37 034	678	$3.8\ 10^{-8}$
a_2		9 139	1	9 139	167	$2.4\ 10^{-6}$
a_{12}		16 056	1	16 056	294	$4.6\ 10^{-7}$
Residue 5		219	4	55		
Total	179 082	179 082	8			

3.2.10 Types de SS

$$SS(A|B) = SS(A,B) - SS(A)$$

Type I (sequential)

$$SS(a_o) \text{ for } a_o$$
$$SS(a_1|a_o) \text{ for } a_1$$
$$SS(a_2|a_o,a_1) \text{ for } a_2$$
$$SS(a_{12}|a_o,a_1,a_2) \text{ for } a_{12}$$

Type II

$$SS(a_o|a_1,a_2,a_{12}) \text{ for } a_o$$
$$SS(a_1|a_o,a_2,a_{12}) \text{ for } a_1$$
$$SS(a_2|a_o,a_1,a_{12}) \text{ for } a_2$$
$$SS(a_{12}|a_o,a_1,a_2) \text{ for } a_{12}$$

3.2.11 Comparison between type I and type II

Source	SS*	DF	MS	F	P
a_1	37 033	1	37 033	678	0.001 %
a_2	9 074	1	9 074	166	0.021 %
a_{12}	16 122	1	16 122	295	0.007 %
Résidu 1	219	4	55		
Total	179 082	8			

Source	SS*	DF	MS	F	P
a_1	7 130	1	7 130	130	0.034 %
a_2	4 700	1	4 700	86	0.075 %
a_{12}	16 122	1	16 122	295	0.007 %
Résidu 1	219	4	55		
Total	179 082	8			

3.3.1 What is an alias ?

- Alias : Zorro and Diego de la Vega
- The Concept of alias is useful for dealing with non-orthogonal situations
- It let compute the connection between two parts of a model
- Examples :
  - For a given design, what is the consequence of considering or not a regressor ?
  - For a given design, what is the consequence of considering the second degree coefficients on the first degree coefficients ?
  - In some situations we will see that the fact of considering the second degree coefficients will change the value of the constant, meaning that the second degree coefficients are aliased with the constant.

3.3.2 The alias matrix

The alias matrix corresponds to the projection of the base vectors of the second subspace on the base vectors of the first sub-space

- Let's consider a linear model $y = f(x_1, \dots, x_N, a_o, a_1, \dots, a_M)$ and a design with the model matrix X
- Now let's part the model in two parts $f = f_1 + f_2$ with he corresponding model matrix X_1 et X_2 so that $X = [X_1, X_2]$
- The alias matrix A of X_2 in relation to X_1 is :

$$A = ({X_1}^T X_1)^{-1} \ ({X_1}^T X_2)$$

3.3.3 What is the alias used to ?

To make the link between the two subspace :

$$\begin{aligned}\hat{Y}_1 &= X_1\hat{\alpha}\\ \hat{Y} &= [X_1\ X_2]\begin{bmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\end{bmatrix}\end{aligned}$$

Coefficients : $\hat{\alpha} = \hat{\alpha}_1 + A\ \hat{\alpha}_2$

Orthogonal projection : $X_{2.1} = X_2 - X_1 A$

Orthogonal decomposition :

$$Y = X_1(\hat{\alpha}_1 + A\ \hat{\alpha}_2) + (X_2 - X_1 A)\alpha_2 + \epsilon$$
$$Y = X_1\alpha + X_{2.1}\alpha_2 + \epsilon$$

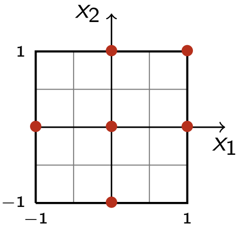
3.3.4 Example of an alias matrix

- Let's consider a design E and the model
  $y = a_o + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$

$$E = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{pmatrix}$$

- the separation is between the linear and the interaction parts :

$$f_1(x) = a_o + a_1 x_1 + a_2 x_2$$
$$f_2(x) = a_{12} x_1 x_2$$



- $X_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{pmatrix}$ et $X_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

3.3.5 Example of an alias matrix (2)

$$A = (X_1^T X_1)^{-1}(X_1^T X_2)$$

$$= \frac{1}{44}\begin{pmatrix} 8 & -2 & -2 \\ -2 & 17 & -5 \\ -2 & -5 & 17 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{pmatrix}^T \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$= \frac{1}{44}\begin{pmatrix} 8 & -2 & -2 \\ -2 & 17 & -5 \\ -2 & -5 & 17 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/11 \\ 5/22 \\ 5/22 \end{pmatrix} \Rightarrow \begin{cases} l_o = a_o + \frac{1}{11}a_{12} \\ l_1 = a_1 + \frac{5}{22}a_{12} \\ l_2 = a_2 + \frac{5}{22}a_{12} \end{cases}$$

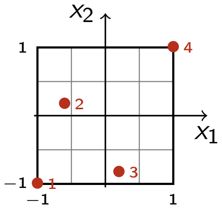
3.3.6 ANOVA table for non orthogonal parts

Source	SS	SS*
X_1	$\alpha_1' X_1' X_1 \alpha_1$	$\alpha' X_1' X_1 \alpha = (\alpha_1 + A\alpha_2)' X_1' X_1 (\alpha_1 + A\alpha_2)$
X_2	$\alpha_2' X_2' X_2 \alpha_2$	$\alpha_2' X_{2.1}' X_{2.1} \alpha_2 = \alpha_2' (X_2 - X_1 A)' (X_2 - X_1 A) \alpha_2$
Résidu	$\epsilon' \epsilon$	
Total	SS_Y	–

N is the number of runs, A the alias matrix relative, α the coefficients of the first part of the model when it is inferred alone and $X_{2.1}$ is the model matrix of the second part of the model orthogonal to the first part.

3.3.7 Let's go back to the case of the cocktail

Runs



Model matrix

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -0.6 & 0.17 & -0.1 \\ 1 & 0.2 & -0.83 & -0.17 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -0.6 & 0.17 & -0.1 \\ 1 & 0.2 & -0.83 & -0.17 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Corrected sum of squares

$a_o|a_1, a_2, a_{12} \rightarrow A_o = (-0.1 \ \ -0.165 \ 0.433)$

which means that :

$a_o^* = a_o - 0.1a_1 - 0.165a_2 + 0.433a_{12}$

Then :

$SS(a_o|a_1, a_2, a_{12}) =$

$$\left(a_o + A_o \begin{bmatrix} a_1 \\ a_2 \\ a_{12} \end{bmatrix}\right)^T X_o^T X_o \left(a_0 + A_o \begin{bmatrix} a_1 \\ a_2 \\ a_{12} \end{bmatrix}\right)$$

And so on for the next steps :

$$a_o, a_1|a_2, a_{12} \rightarrow A_1 = \begin{pmatrix} -0.09 & 0.44 \\ 0.71 & 0.09 \end{pmatrix}$$

$$a_o, a_1|a_2, a_{12} \rightarrow A_2 = \begin{pmatrix} 0.46 \\ -0.04 \\ 0.18 \end{pmatrix}$$

Note : Matrices and arrays in Matlab

- Arrays are the **fundamental data type** used to store collections of data in the form of elements arranged in rows and columns.
- Arrays can be one-dimensional (vectors) or two-dimensional (matrices), but MATLAB also supports **multidimensional** arrays.
- Arrays can hold **various types of data**, such as numbers, strings, or even more complex objects.
- Most operations in MATLAB are **vectorized**, meaning that they are applied element-wise to arrays, which makes computations with arrays fast and efficient.
- The function *repmat(X,v,l)* create a new array by repetition of X, v times vertically and l times horizontally.

```
>> V=[1,2;3,4]

V =

     1     2
     3     4

>> U=ones(3)

U =

     1     1     1
     1     1     1
     1     1     1

>> I=eye(3)

I =

     1     0     0
     0     1     0
     0     0     1
```

Note : cell array

- **Cell arrays** are a type of data structure that allows you to store elements of varying types and sizes.
- Unlike regular arrays, a **cell array can hold different types of data in each of its cells.**
- Each element in a cell array is accessed using **curly braces { }** to retrieve the actual content inside the cell.

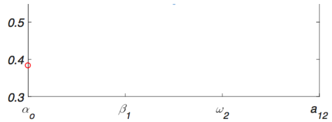
```
>> label={'\alpha_o' '\beta_1' '\omega_2' 'a_{12}'}

label =

  1×4 cell array

    '\alpha_o'    '\beta_1'    '\omega_2'    'a_{12}'

>> plot(1:4,sin([1:4]*pi/8),'or')
>> set(gca,'XTick',1:4,'XTickLabel',label)
>> |
```



Note : Table in Matlab

- A table is a data type specifically designed to store and organize heterogeneous data, where **each column can hold a different type of data** (e.g., numerical, text, or categorical).

```
Nom={'Alluminium';'Plomb';'Cuivre';'Fer'};
E=[0.72E11;0.17E11;1.00E11;2.20E11];
mu=[.34;.45;.34;.30];
Symbole={'Al';'Pb';'Cu';'Fe'};

T=table(E,mu,Symbole,'RowNames',Nom);
```

	E	mu	Symbole
Alluminium	7.2e+10	0.34	'Al'
Plomb	1.7e+10	0.45	'Pb'
Cuivre	1e+11	0.34	'Cu'
Fer	2.2e+11	0.3	'Fe'

Note : Linear model in Matlab

Linearmodel is an object created by routines such as *fitlm* or *stepwiselm* and with the following content

- experimental data,
- model description,
- statistics for a diagnostic,
- estimated coefficients,
- residuals.

The object can be reused to predict the responses of the model with the methods *predict* and *feval*

3.3.8 Routines *fitlm* and *stepwiselm*

These MATLAB functions return a linear regression model fit to variables in the table or dataset array.

Matlab

- *mdl=fitlm(tbl)*
  mdl=fitlm(tbl,modelspec)
  mdl=fitlm(x,y)
  mdl=fitlm(x,y,modelspec)
  mdl=fitlm(...,Name,Value)
- *mdl=stepwiselm(tbl,modelspec)*
  mdl=stepwiselm(x,y,modelspec)
  mdl=stepwiselm(...,Name,Value)

These routines can be fed by tables or arrays

3.3.9 *fitlm( )* output

```
Linear regression model:
    R ~ 1 + Var1*Var2

Estimated Coefficients:
                   Estimate      SE      tStat       pValue
                   ________    ______    ______    __________

    (Intercept)     97.643     3.4385    28.397     9.1511e-06
    Var1              52.5     4.5959    11.423     0.00033506
    Var2            41.297     4.4527    9.2748     0.00075164
    Var1:Var2       81.214     4.7279    17.178     6.7382e-05


Number of observations: 8, Error degrees of freedom: 4
Root Mean Squared Error: 7.39
R-squared: 0.997,  Adjusted R-Squared 0.994
F-statistic vs. constant model: 380, p-value = 2.29e-05
```

3.3.10 Wilkinson's notation

Wilkinson notation is a concise way to specify the terms in a linear model. It describes the relationships between predictors (independent variables) and a response (dependent variable) without explicitly stating the coefficients of the model.

Example : $Y \sim 1 + X_1 * X_2$ represents the model $y = a_o + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$

Termse of the model	Wilkinson's notation
intercept a_o	1
sans intercept	-1
a_1	$X1$
a_1, a_2	$X1 + X2$
a_1, a_2, a_{12}	$X1 * X2$ ou $X1 + X2 + X1 : X2$
a_{12}	$X1 : X2$
a_1, a_{11}	$X1^2$
a_{11}	$X1^2 - X1$

3.3.11 Methods for linear model objects

Matlab offers several methods to be used with *linearmodel* objects

Matlab

- *anova(mdl)*
- *coefCI(mdl)*
- *coefTest(mdl), coefTest(mdl,H,C)*
- *plot(mdl)*
- *plotAdded(mdl,coef)*
- *plotDiagnostics(mdl, plottype)*
- *plotResiduals(mdl)*
- *plotEffects(mdl)*
- *ypred = predict(mdl,Xnew)*

1. Significance of Coefficients and R-squared

- **Check p-values :**
  - Small p-values (< 0.05) indicate that the corresponding predictors are statistically significant.
- **R-squared :**
  - Measures the proportion of variance explained by the model.
  - Ranges from 0 to 1 ; higher values indicate a better fit.
- **Adjusted R-squared :**
  - Adjusts for the number of predictors in the model.
  - Prevents overfitting by penalizing for additional, non-informative variables.

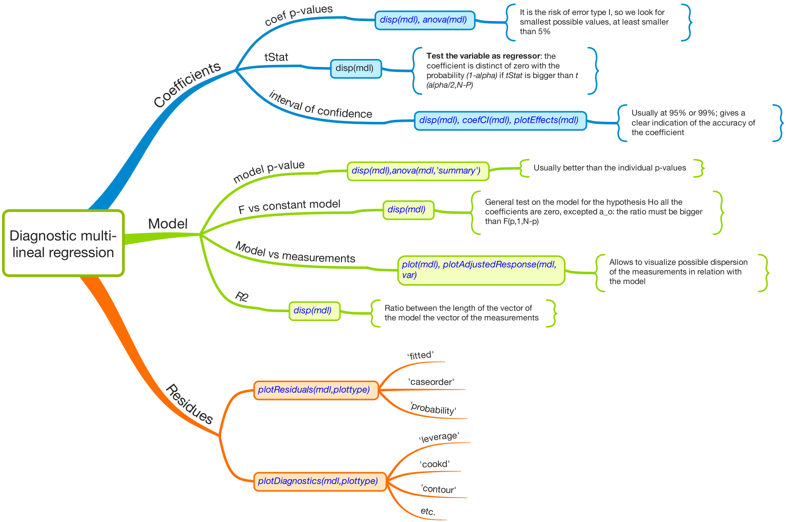
2. Residual Analysis and Normality of Residuals

- **Residual Plot :**
  - Check if residuals are randomly scattered around zero.
  - Patterns (curvature, funnel shape) may indicate misspecification.
- **Normality of Residuals :**
  - Use a Q-Q plot to check if residuals follow a normal distribution.
  - Apply the Shapiro-Wilk test for a formal test of normality.

3. Homoscedasticity and Leverage & Influence

- **Homoscedasticity (Constant Variance) :**
  - Residuals should have constant variance across the range of predictors.
  - Breusch-Pagan test can formally check for heteroscedasticity.
- **Leverage and Influence :**
  - Use leverage statistics to detect points with large influence on the model.
  - Cook's distance can help identify outliers that may disproportionately affect the model.

3.3.12 Diagnostic of a LSF



3.3.13 *candexch* routine

The *candexch()* function in MATLAB is used to generate optimal experimental designs. It is commonly used when working with a set of candidate points to select the most informative subset for fitting a model.

It selects a subset of points from a candidate set that maximizes the D-optimality criterion, ensuring the most information is gained from the least number of experimental runs. By exchanging points iteratively, it refines the design to provide a robust and efficient design for fitting statistical models.

Matlab

- *list=candexch(X,nrows)*
- This routine *candidate exchange* allows the selection of the best *nrows* of a model matrix in the D-optimal perspective,
- The standard routine proposes duplicated points,

3.3.14 Summary : ANOVA in DOE

- **Purpose of ANOVA :**
  - Decomposes total variance into components (Model and Residual).
  - Tests the significance of factor effects and interactions.
- **Key Concepts :**
  - **Sum of Squares (SS) :** Measures variability attributed to factors.
  - **Mean Squares (MS) :** SS divided by degrees of freedom (DF).
  - **F-statistic :** Ratio of MS for model terms to MS for residuals.
  - **p-values :** Indicates significance of factors.
- **Model Interpretation :**
  - Significant terms (p-value < 0.05) indicate meaningful effects.
  - Main effects and interactions are analyzed through ANOVA tables.
- **Common Applications :**
  - Factorial experiments : Assess main effects and interactions.
  - Response surface methodology : Investigate curvature and optimization.