Chapitre 4 Analyse statistique

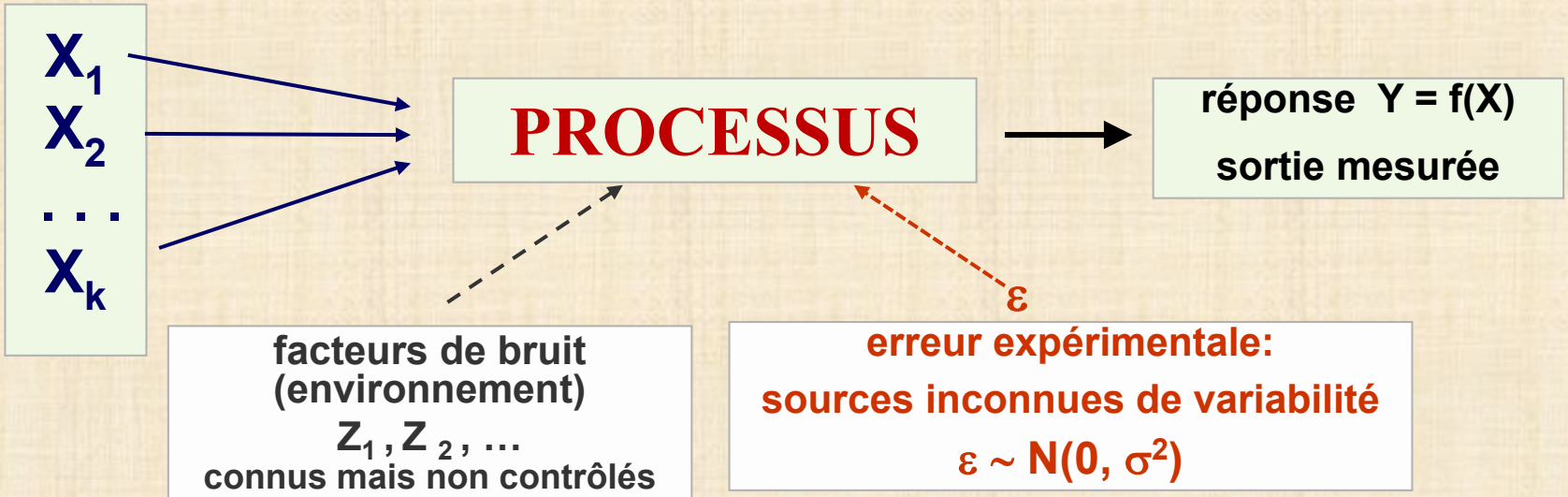
- MODÈLES STATISTIQUES
- ANALYSE STATISTIQUE : étapes
- RÉGRESSION MULTIPLE
  - ajustement du modèle
  - analyse de la variabilité - ANOVA
  - analyse des résidus
  - qualité d'un bon modèle
  - utilisation de *STATISTICA*
- ANALYSE du PLAN 16 ESSAIS
  - modèle avec 4 facteurs
  - modèle avec 5 facteurs
  - modèle avec 8 facteurs
  - étapes de l'analyse
  - calculs

MODÉLISATION STATISTIQUE

Toute analyse statistique repose sur un MODÈLE

- fonction f pour représenter une relation entre input X et output Y
- hypothèse distributionnelle pour le terme d'erreur ε

$X_1, X_2, \ldots, X_k$: facteurs contrôlés (expérimentation)
facteurs mesurés (mode passif)



$$Y = f(X_1, X_2, \ldots, X_k ; \beta_0, \beta_1, \beta_2, \ldots) + \varepsilon$$

f : fonction inconnue → approximation polynôme

$\beta_0, \beta_1, \beta_2, \ldots$: paramètres statistiques inconnus

MODÈLES : types

- **effets principaux (ordre 1) :** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ (1)
- **effets principaux et interaction :**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \ldots \quad (2)$$

- **quadratiques (facteurs quantitatifs) : ordre 2**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \ldots +$$

$$+ \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \ldots \quad (3)$$

- **d'analyse de la variance : variables catégoriques** (4)

  transformées en plusieurs variables indicatrices $X = 0 / 1$

  ou en plusieurs variables de type $X = -1 / 0 / 1$ **(codage à effet)**

- **mixtes : facteurs fixes + facteurs aléatoires** (5)
- **polynomial :** $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \bullet\ \bullet\ \bullet + \beta_k X^k$ **peu utilisé**

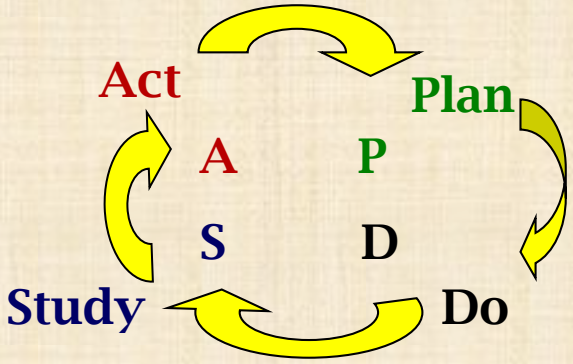
EXPÉRIMENTATION : étapes

PHASE	ÉTAPES
P: planifi – cation	1 Définir PROCESSUS / problématique / objectifs
	2 Choisir les variables de RÉPONSE (S) Y à mesurer
	3 Choisir les VARIABLES facteurs X et l'espace de variation
	4 Choisir et comparer des PLANS EXPÉRIMENTAUX
D: exécution	5 PRÉPARER pour l'expérience
	6 CONDUIRE l'expérience
S: analyse	7 ANALYSE statistique des résultats
A: transfert	8 AGIR avec les conclusions de l'analyse

analyse ? ch5

roue PDSA

Shewhart - Deming



ANALYSE STATISTIQUE : processus (1/2)

Étape 7

7.1 Spécification d'un modèle statistique

7.2 Estimation des paramètres du modèle

7.3 Décomposition de la variabilité : ANOVA

7.4 Tests d'hypothèses sur les paramètres

7.5 Analyse diagnostique des résidus

- vérification des hypothèses de base
- identification d'observations influentes
- transformation Box-Cox de réponse Y

7.6 Si nécessaire : itération des étapes 1 à 5

7.7 Optimisation de la réponse (s'il y a lieu)

7.8 Graphiques de la réponse

ANALYSE STATISTIQUE : processus (2/2)

1. Préparation de la matrice de tests pour la collecte des données
2. Spécification modèle pour l'analyse

   MATRICE MODÈLE = MATRICE DESIGN + COLONNES ADDITIONNELLES

   (augmentée) = (effets principaux) + (effets d'interaction)

3. ***Optionnel*** : examen de la variable de réponse avec des cartes de Shewhart (si on a des répétitions)
4. Ajustement du modèle : estimation des paramètres statistiques β_0 , β_1 , …
5. Calcul du tableau d'analyse de la variance : ANOVA
6. Tests d'hypothèses des paramètres (effets) : β_1, β_2 , …
7. Analyse diagnostique des résidus
8. ***Optionnel*** : Itération des étapes 4-5-6-7 - modèle avec effets importants seulement
9. Présentation graphiques des résultats
   - diagramme Pareto
   - diagrammes effets principaux
   - diagrammes interactions
   - courbes contour

RÉGRESSION LINÉAIRE MULTIPLE (1 / 17)

MODÈLE $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \bullet\bullet\bullet + \beta_m X_m + \varepsilon$ $\qquad \varepsilon \sim N(0, \sigma^2)$

DONNÉES

#	X_0	X_1	X_2	X_3	.....	X_m		Y
1	1	x_{11}	x_{12}	x_{13}	.....	x_{1m}		y_1
2	1	x_{21}	x_{22}	x_{23}	.....	x_{2m}		y_2
.	.	.	.	.	.....	.		.
i	1	x_{i1}	x_{i2}	x_{i3}	.....	x_{im}		y_i
.	.	.	.	.	.....	.		.
N	1	x_{N1}	x_{N2}	x_{N3}	....	x_{Nm}		y_N

$X_{i0} = 1$: effet général

matrice des effets : colonnes types X_j , $X_j X_k$, X_j^2

écriture matricielle $X_{N \times p} = [\, x_{ij} \,]$ $\qquad p = 1 + m$ $\qquad Y_{N \times 1}$: vecteur N x 1

$\beta_{p \times 1} = (\beta_0 , \beta_1 , \beta_2 , \beta_3 , ... , \beta_k)'$: vecteur p x 1

remarque : l'opération de transposition de vecteurs / matrices est notée par le symbole '

ESTIMATION (principe de moindres carrés) : $\underset{\beta}{\text{Min}} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2$

système d'équations linéaire à résoudre : $(X' X)\, \beta = X' Y$

solution : $\hat{\beta} = (X' X)^{-1} X' Y = C\, Y$

$C = (X' X)^{-1} X'$ est une matrice p x N de valeurs fixes connues

ÉQUATION de PRÉDICTION $\hat{y} = X \hat{\beta}$

RÉGRESSION LINÉAIRE MULTIPLE (2 / 17)

PROPRIÉTÉS ESTIMATEURS $\hat{\beta}$

- combinaisons linéaires des y_i
- sans biais : $E(\hat{\beta}_j) = \beta_j$ pas d'erreur systématique
- $var(\hat{\beta}) = (X' X)^{-1} \sigma^2$ variance minimale (« meilleurs »)

ESTIMATION de σ^2

résidu : $e_i = \hat{y}_i - y_i$

somme de carrés résiduels : $SSR = \sum e_i^2$

carré résiduel moyen : $MSR = SSR / N - m - 1$

estimation : $\hat{\sigma}^2 = MSR$ $\hat{\sigma} = (MSR)^{0.5}$

DÉCOMPOSITION DE LA VARIABILITÉ : tableau d'analyse de la variance

$SSY = \sum (y_i - \bar{y}.)^2$: somme TOTALE des carrés

$SSM = \sum (\hat{y}_i - \bar{y}.)^2$: somme des carrés du MODÈLE (régression)

$SSR = \sum (\hat{y}_i - y_i)^2$: somme des carrés RÉSIDUELS

ÉQUATION FONDAMENTALE

somme de carrés (SS) : SSY = SSM + SSR

variabilité : totale = modèle + résiduelle

degrés de liberté (df) : N – 1 = m + (N – m - 1)

RÉGRESSION LINÉAIRE MULTIPLE (3 / 17)

TABLEAU D'ANALYSE VARIANCE : modèle de régression linéaire multiple

SOURCE	df	SS	MS=SS / df	F-ratio	p-valeur
régression	m	SSM	MSM = SSM / m	f = MSM / MSR	P(F ≥ f)
résiduelle	N – m - 1	SSR	MSR = SSR /(N – m - 1) = $\hat{\sigma}^2$	---------	-------
totale	N – 1	SSY	----------	--------	-------

R^2 = SSM / SSY : coefficient de détermination

$0 \leq R^2 \leq 1$: fraction de la variabilité de Y expliquée par les variables X

$R^2_{adj} = 1 - [(N - 1)/(N - m)] (1 - R^2)$: coefficient de détermination ajusté

remarques

- ajouter une variable explicative additionnelle dans un modèle augmente SSM et R^2
- l'augmentation de R^2 n'est pas toujours importante et significative
- R^2_{adj} est préférable à R^2 pour comparer deux modèles

Test global $H_{0G} : \beta_1 = \beta_2 = \ldots = \beta_m = 0$

vs H_{1G} : non H_{0G} (au moins un $\beta \neq 0$)

rejeter H_0 au seuil α si $f > F_{m,\, N-m-1,\, 1-\alpha}$

RÉGRESSION LINÉAIRE MULTIPLE (4 / 17)

Distribution d'échantillonnage de $\hat{\beta}_j$ $\quad j = 0, 1, 2, \ldots, k$

$$[(\hat{\beta}_j - \beta_j) / \hat{\sigma} (c_{jj})^{0.5}] \sim t_{N-m-1} \qquad c_{jj} : \text{j-ème élément diagonal de } (X' X)^{-1}$$

Applications

(a) test $H_{0j} : \beta_j = 0$ vs $H_{1j} : \beta_j \neq 0 \qquad j = 1, 2, \ldots, m$

rejeter H_{0j} au seuil α si $\quad | \hat{\beta}_j | / \hat{\sigma} (c_{jj})^{0.5} > t_{N-m-1,\, 1-\alpha/2}$

(b) Intervalle de confiance $\quad \beta_j : \hat{\beta}_j \pm t_{N-m-1,\, 1-\alpha/2} \, \hat{\sigma} (c_{jj})^{0.5}$

(c) INTERVALLE de CONFIANCE MOYENNE de Y à $X_1 = x_1^*, X_2 = x_2^*, \ldots, X_k = x_k^*$

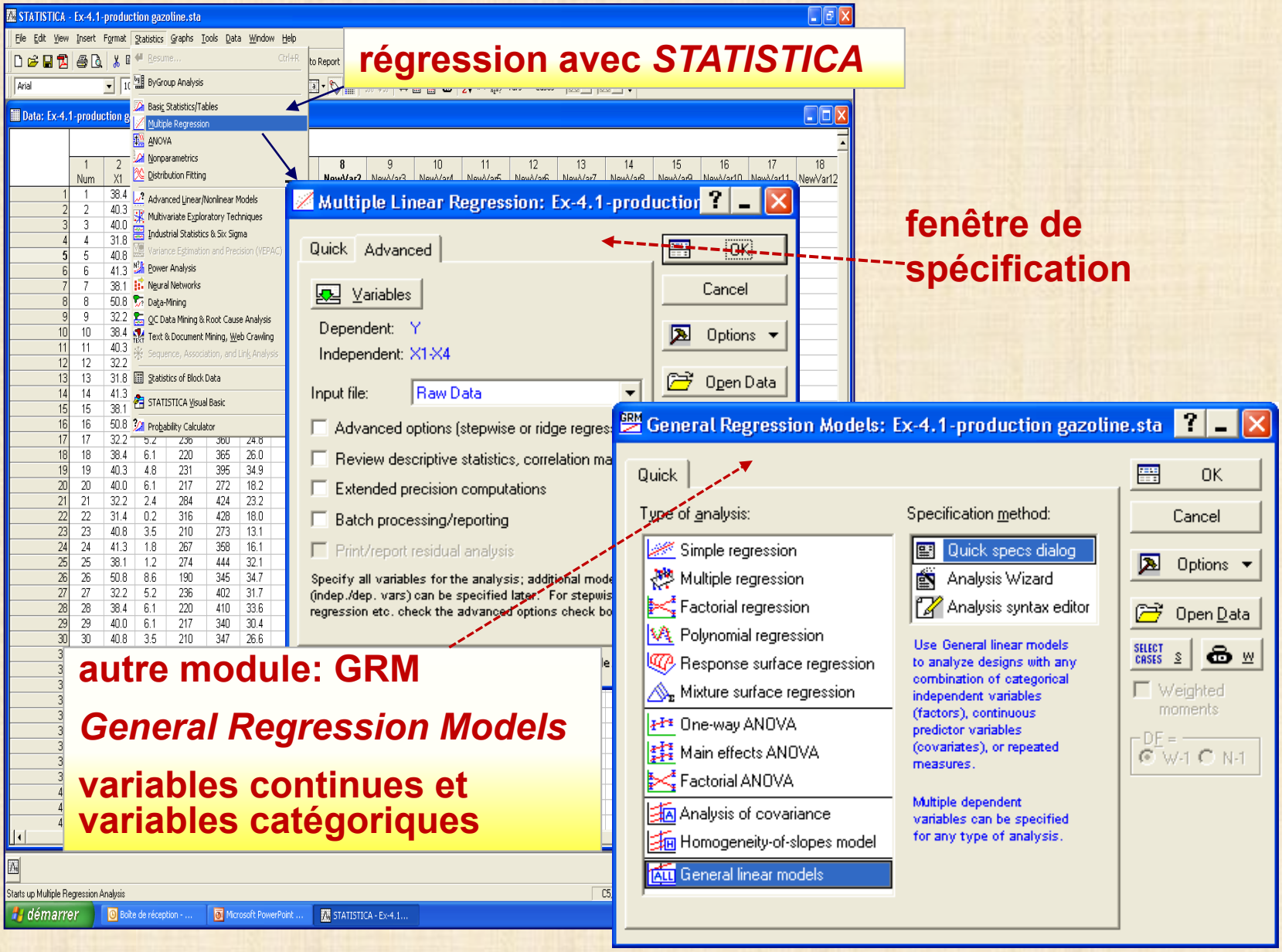
$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^* \qquad x^* = (x_1^*, x_2^*, \ldots, x_k^*)$$

$$E (Y \mid X = x^*) : \quad \hat{y}^* \pm t_{N-m-1,\, 1-\alpha/2} \, \hat{\sigma} [x^* (X' X)^{-1} x^{*\prime}]^{0.5}$$

(d) INTERVALLE de PRÉDICTION VALEUR de Y à $X_1 = x_1^*, X_2 = x_2^*, \ldots, X_k = x_k^*$

$$Y \mid X = x^* : \quad \hat{y}^* \pm t_{N-m-1,\, 1-\alpha/2} \, \hat{\sigma} [1 + x^* (X' X)^{-1} x^{*\prime}]^{0.5}$$

RÉGRESSION LINÉAIRE MULTIPLE (5 / 17)



RÉGRESSION LINÉAIRE MULTIPLE (6 / 17)

Exemple: production de gazoline avec huiles brutes (données historiques)

N. H. Prater, *Petroleum Refiner - Experimental Designs in Industry* (ed. Chew) Wiley 1956 pp 109-137

Y : rendement production gazoline (% de l'huile brute)

X1 : gravité huile brute (deg. API) X2 : pression vapeur (PSIA)

X3 : ASTM point 10% (deg. F) X4 : point sortie gazoline (deg. F)

#	X1	X2	X3	X4	Y
1	38.4	6.1	220	235	6.9
2	40.3	4.8	231	307	14.4
3	40.0	6.1	217	212	7.4
4	31.8	0.2	316	365	8.5
5	40.8	3.5	210	218	8.0
6	41.3	1.8	267	235	2.8
7	38.1	1.2	274	285	5.0
8	50.8	8.6	190	205	12.2
9	32.2	5.2	236	267	10.0
10	38.4	6.1	220	300	15.2
11	40.3	4.8	231	367	26.8
12	32.2	2.4	284	351	14.0
13	31.8	0.2	316	379	14.7
14	41.3	1.8	267	275	6.4
15	38.1	1.2	274	365	17.6
16	50.8	8.6	190	275	22.3
17	32.2	5.2	236	360	24.8
18	38.4	6.1	220	365	26.0
19	40.3	4.8	231	395	34.9
20	40.0	6.1	217	272	18.2
21	32.2	2.4	284	424	23.2
22	31.4	0.2	316	428	18.0
23	40.8	3.5	210	273	13.1
24	41.3	1.8	267	358	16.1
25	38.1	1.2	274	444	32.1
26	50.8	8.6	190	345	34.7
27	32.2	5.2	236	402	31.7
28	38.4	6.1	220	410	33.6
29	40.0	6.1	217	340	30.4
30	40.8	3.5	210	347	26.6
31	41.3	1.8	267	416	27.8
32	50.8	8.6	190	407	45.7

Y a t-il une structure dans ces données?

réponse: oui

RÉGRESSION LINÉAIRE MULTIPLE (7 / 17)

#	X1	X2	X3	groupe	X4	Y
8	50,8	8,6	190	1	205	12,2
16	50,8	8,6	190	1	275	22,3
26	50,8	8,6	190	1	345	34,7
32	50,8	8,6	190	1	407	45,7
6	41,3	1,8	267	2	235	2,8
14	41,3	1,8	267	2	275	6,4
24	41,3	1,8	267	2	358	16,1
31	41,3	1,8	267	2	416	27,8
5	40,8	3,5	210	3	218	8,0
23	40,8	3,5	210	3	273	13,1
30	40,8	3,5	210	3	347	26,6
2	40,3	4,8	231	4	307	14,4
11	40,3	4,8	231	4	367	26,8
19	40,3	4,8	231	4	395	34,9
3	40,0	6,1	217	5	212	7,4
20	40,0	6,1	217	5	272	18,2
29	40,0	6,1	217	5	340	30,4
1	38,4	6,1	220	6	235	6,9
10	38,4	6,1	220	6	300	15,2
18	38,4	6,1	220	6	365	26,0
28	38,4	6,1	220	6	410	33,6
7	38,1	1,2	274	7	285	5,0
15	38,1	1,2	274	7	365	17,6
25	38,1	1,2	274	7	444	32,1
9	32,2	5,2	236	8	267	10,0
17	32,2	5,2	236	8	360	24,8
27	32,2	5,2	236	8	402	31,7
12	32,2	2,4	284	9	351	14,0
21	32,2	2,4	284	9	424	23,2
4	31,8	0,2	316	10	365	8,5
13	31,8	0,2	316	10	379	14,7
22	31,4	0,2	316	10	428	18,0

données en structure emboitée

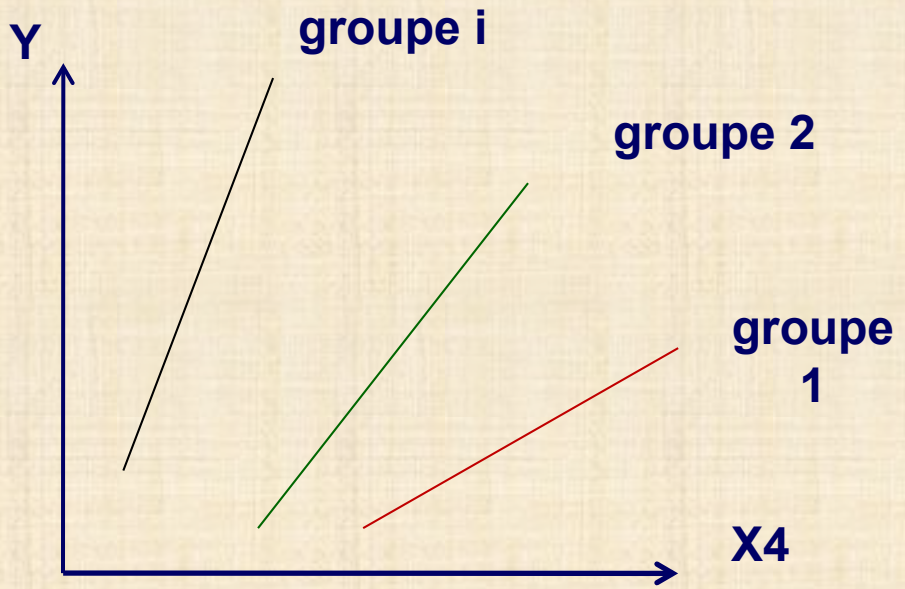
10 groupes d'huile brute définis par X1 X2 X3

régression de Y sur X4

1 modèle pour chaque groupe

analyse de covariance =

variables continues + variables catégoriques

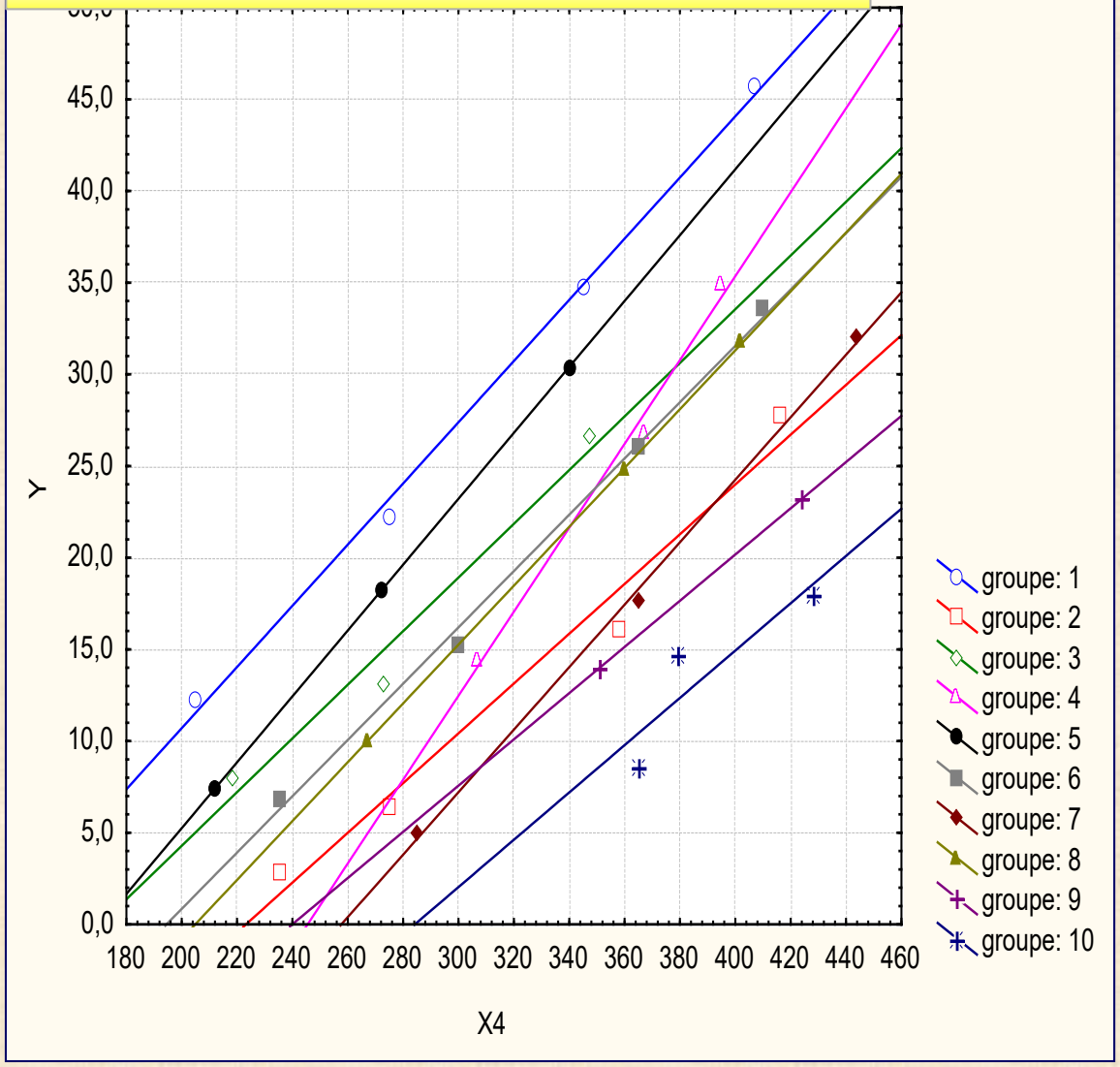


modèle 1: pentes distinctes

modèle 2: pente égales – ANCOVA

RÉGRESSION LINÉAIRE MULTIPLE (8 / 17)

Modèle : pente égales – ANCOVA



groupe: 1 Y = -22,67 + 0,167*x

r = 0,999; p = 0,0012; r^2 = 0,9976

groupe: 2 Y = -30,29 + 0,136*x

r = 0,9875; p = 0,0125; r^2 = 0,9752

groupe: 3 Y = -24,97 + 0,146*x

r = 0,9855; p = 0,1084; r^2 = 0,9713

groupe: 4 Y = -56,16 + 0,229*x

r = 0,9963; p = 0,0550; r^2 = 0,9925

groupe: 5 Y = -30,69 + 0,180*x

r = 1,0000; p = 0,0006; r^2 = 1,0000

groupe: 6 Y = -29,86 + 0,153*x

r = 0,9979; p = 0,0021; r^2 = 0,9958

groupe: 7 Y = -43,91 + 0,170*x

r = 0,9990; p = 0,0281; r^2 = 0,9981

groupe: 8 Y = -32,88 + 0,160*x

r = 1,0000; p = 0,0048; r^2 = 0,9999

groupe: 9 Y = -30,24 + 0,126*x

r = 1,0000; p = ---; r^2 = 1,0000

groupe: 10 Y = -36,66 +0,129*x

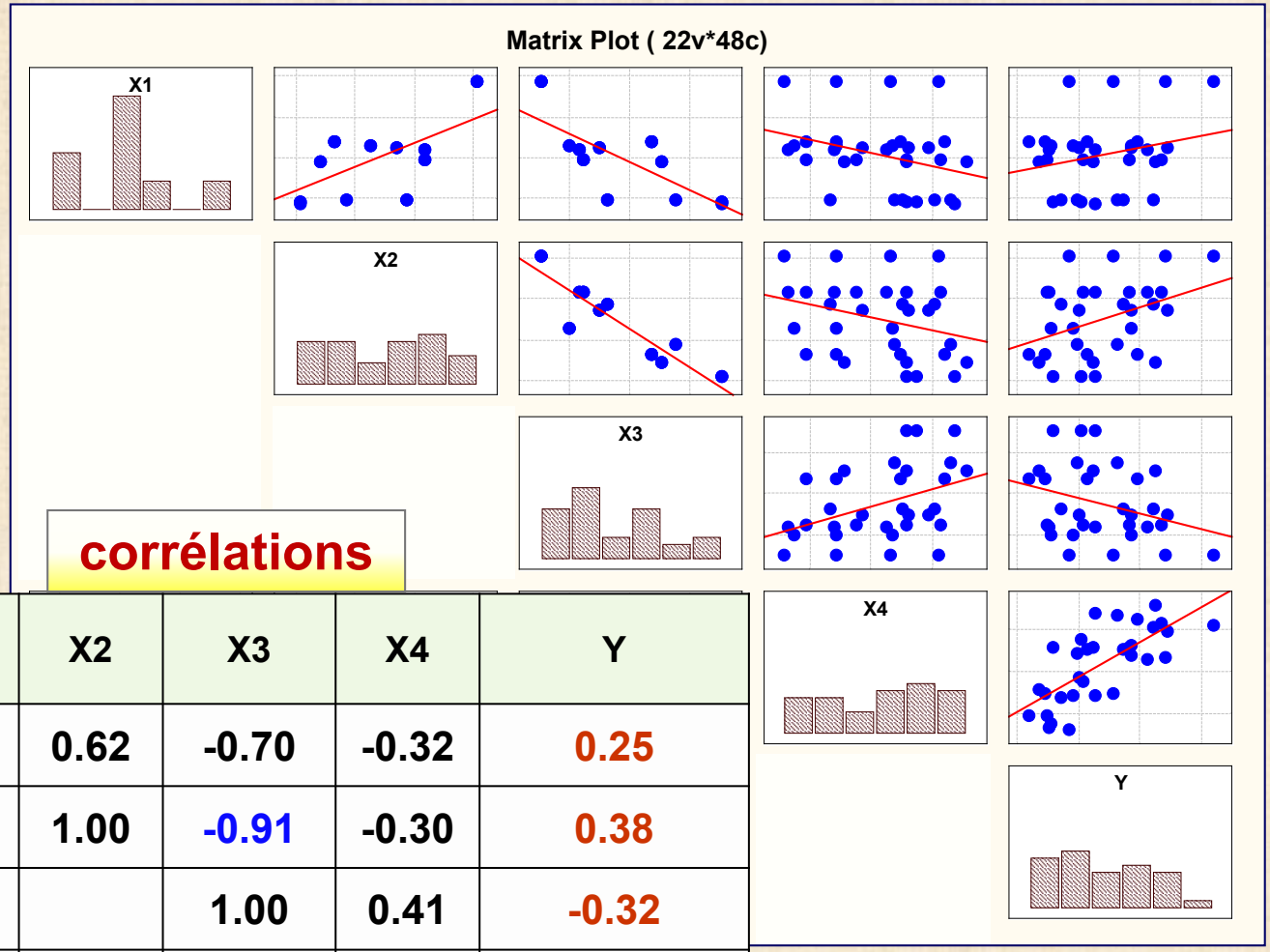
r = 0,8848; p =0,0387; r^2 = 0,7828

RÉGRESSION LINÉAIRE MULTIPLE (9/ 17)

Analyse

Régression

Multiple



corrélations

	X1	X2	X3	X4	Y
X1	1.00	0.62	-0.70	-0.32	0.25
X2		1.00	-0.91	-0.30	0.38
X3			1.00	0.41	-0.32
X4				1.00	0.71
Y					1.00

RÉGRESSION LINÉAIRE MULTIPLE (10/17)

Regression Summary for Dependent Variable: Y
R = 0.981 R^2 = 0.962 Adjusted R^2 = 0.957
F(4,27)=172.06 p < 0.00000 Std.Error of estimate: 2.23

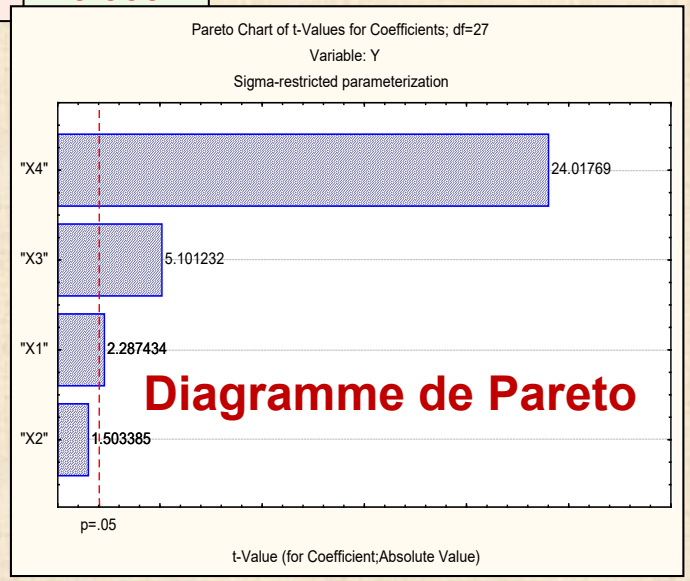
test de Student

variables significatives

	b*	Std.Err. b*	b	Std.Err. b	t(27)	p-level
Intercept			- 6.97	10.13	- 0.69	0.497
X1	0.120	0.053	0.23	0.10	2.29	0.030
X2	0.136	0.090	0.56	0.37	1.50	0.144
X3	- 0.522	0.102	- 0.15	0.03	- 5.10	0.000
X4	1.006	0.042	0.15	0.01	24.02	0.000

coefficients de l'équation de prédiction avec toutes les variables X et Y centrées réduites: permet de mieux comparer la contribution relative des X sur Y : corrélation partielle X_i sur Y avec X_j constants $j \neq i$

coefficients bruts de l'équation de prédiction avec toutes les variables X et Y dans leurs unités d'origine



RÉGRESSION LINÉAIRE MULTIPLE (11 / 17)

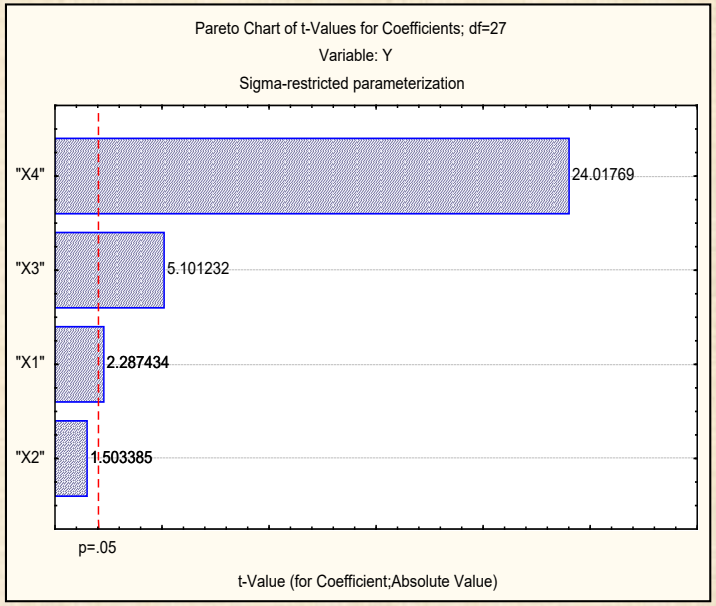
avec module GRM

Regression Summary for Dependent Variable: Y
R = 0.981 R^2 = 0.962 Adjusted R^2 = 0.957
F(4,27)=172.06 p < 0.00000 Std. Error of estimate: 2.23 = $\hat{\sigma}$

	Sums of Square SS	df	Mean Square MS	F	p-level
Regress.	3429.53	4	857.38	172.055	0.000000
Residual	134.55	27	4.98		
Total	3564.08				

	Degr. of freedom	Y SS	Y MS	Y F	Y p
Intercept	1	2.357	2.357	0.4730	0.497485
"X1"	1	26.074	26.074	5.2324	0.030230
"X2"	1	11.263	11.263	2.2602	0.144348
"X3"	1	129.675	129.675	26.0226	0.000023
"X4"	1	2874.54	2874.54	576.849	0.000000
Error	27	134.55	4.983		
Total	31	3564.08			

Diagramme de Pareto



RÉGRESSION LINÉAIRE MULTIPLE (12 / 17)

observations triées ordre résidu	Valeur Obs Y	Valeur Préd $\hat{Y}$	Résidu r $Y - \hat{Y}$	Valeur prédite en centrée réduite	Résidu standard r / sigma	Ecart type Préd ET ($\hat{Y}$)
29 . . . . . *	30.40	25.78	4.62	0.58	2.07	0.540
10 .*	15.20	18.78	-3.58	-0.08	-1.60	0.603
2 .*	14.40	17.93	-3.53	-0.16	-1.58	0.424
19 . . . . . * .	34.90	31.54	3.36	1.13	1.50	0.617
20 . . . . . *. .	18.20	15.26	2.94	-0.42	1.32	0.603
24 . *	16.10	19.01	-2.91	-0.06	-1.30	0.731
. . .	. . .	. . .	. . .	. . .	. . .	. . .
12 . . . *. . . .	14.00	13.65	0.35	-0.57	0.16	0.744
26 . . . *. . . .	34.70	34.43	0.27	1.40	0.12	1.054
14 . . . *. . . .	6.40	6.17	0.23	-1.28	0.10	0.833
31 . . . * . . .	27.80	27.98	-0.18	0.79	-0.08	0.872
30 . . . *	26.60	26.64	-0.04	0.66	-0.02	1.202
Minimum .*	2.80	-0.01	-3.58	-1.87	-1.60	0.424
Maximum .. . *	45.70	44.02	4.62	2.32	2.07	1.223
Moyenne .. * .	19.66	19.66	0.00	0.00	0.00	0.857

analyse des résidus ('model checking')

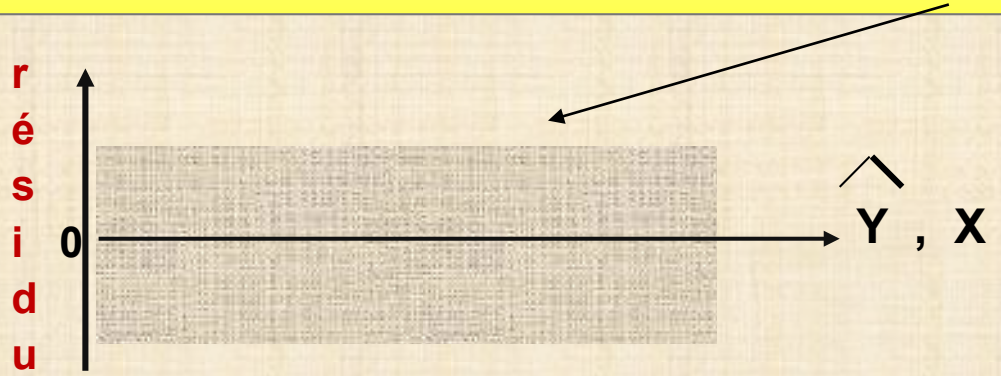
- important de faire une vérification a posteriori quand on ajuste tout modèle statistique
- hypothèses de base
  - variance constante?
  - distribution normale résidus?
  - « bon » modèle? (page suivante)
  - indépendance observations?
  - données aberrantes (ouliers)?
- Si hypothèses de base violées
  - quoi faire ?
  - solution : transformer Y

    transformation de Box-Cox Y^{λ}

    $-2 < \lambda < 2$

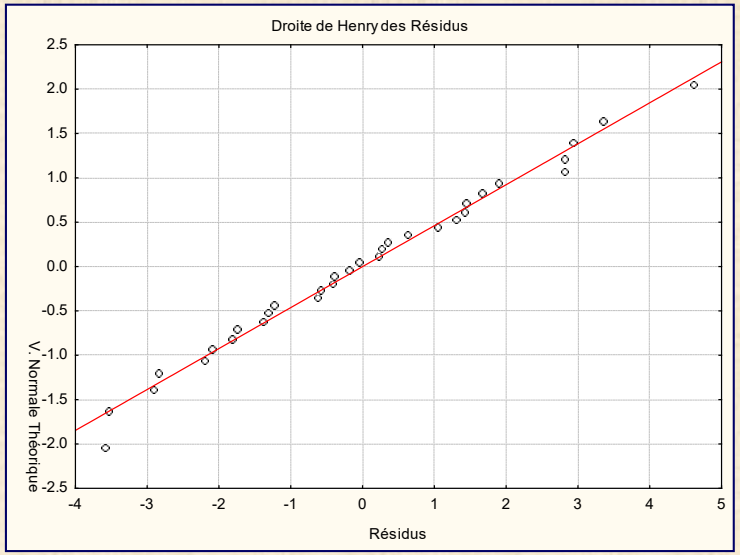
critères «bon modèle» de régression multiple

- test global F significatif
- tests individuels significatifs pour chacun des coefficients du modèle ajusté
- R^2 élevé (au moins 0.70) et R^2_{adj} légèrement inférieur à R^2
- analyse de sensibilité : pas d'observations avec une influence prépondérante
- absence de colinéarité forte entre les variables X
- analyse des résidus ne présente pas d'anomalies
  - indépendance des observations de Y (toujours vrai avec des données expér.)
  - distribution gaussienne : alignement sur un q-q plot (surtout données aberrantes)
  - variance de Y constante
  - graphiques des résidus avec (y observés , chaque X) : bande horizontale

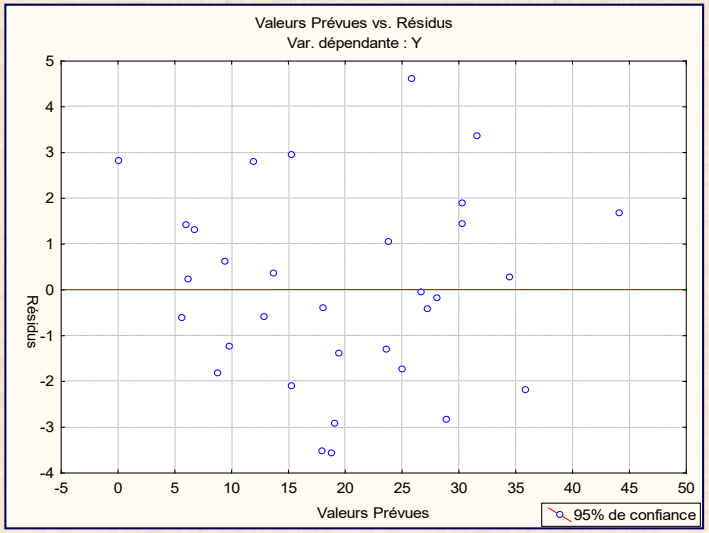


RÉGRESSION LINÉAIRE MULTIPLE (15 / 17)

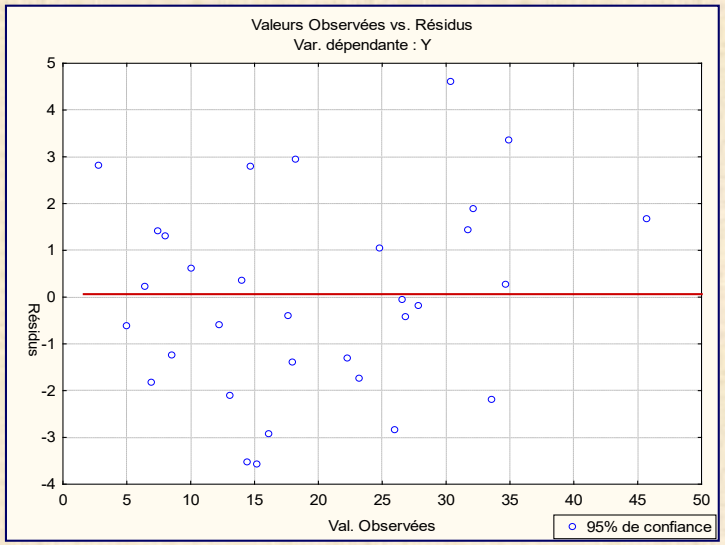
résidus sur échelle gaussienne



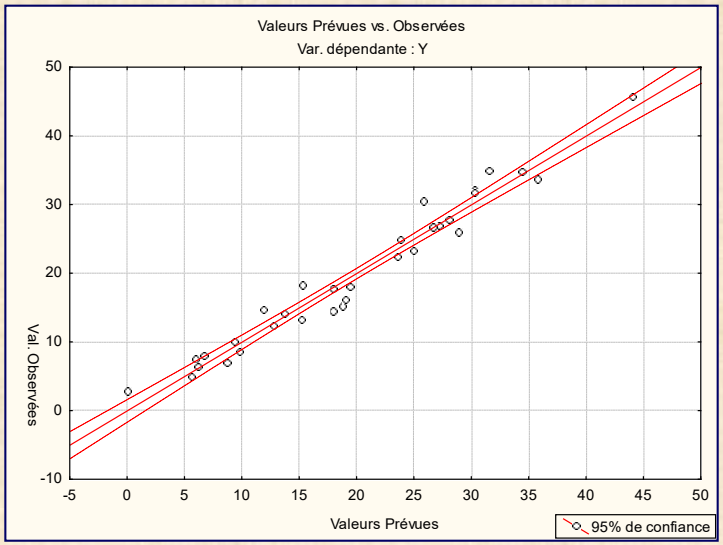
résidus vs prédictions



résidus vs observées

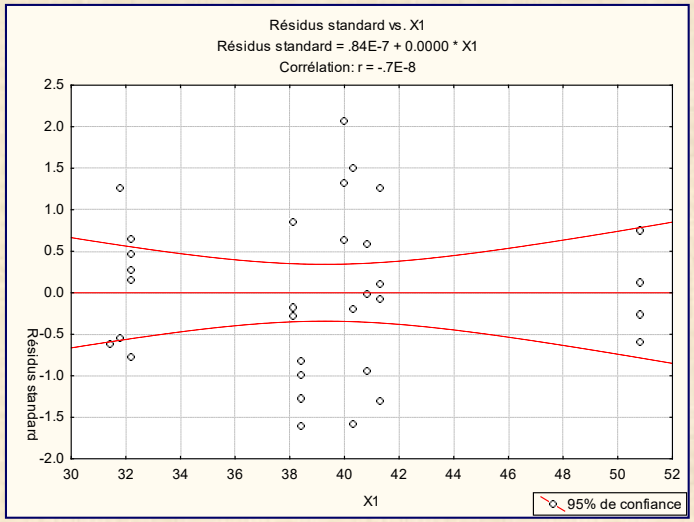


prédictions vs observées

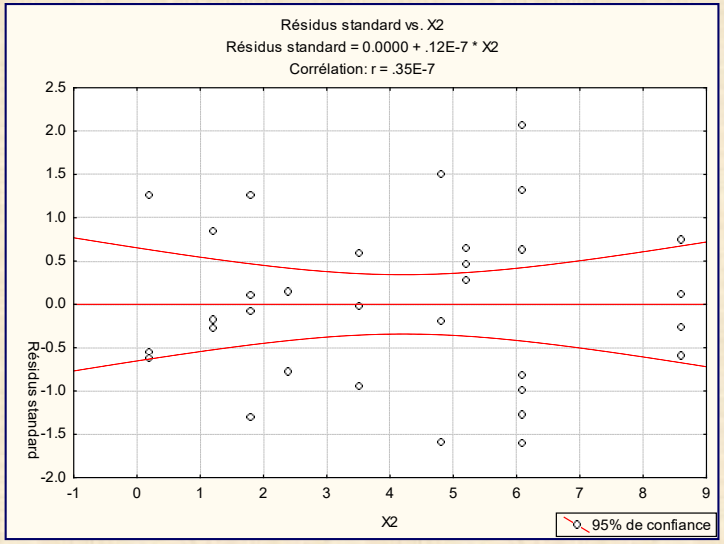


RÉGRESSION LINÉAIRE MULTIPLE (16 / 17)

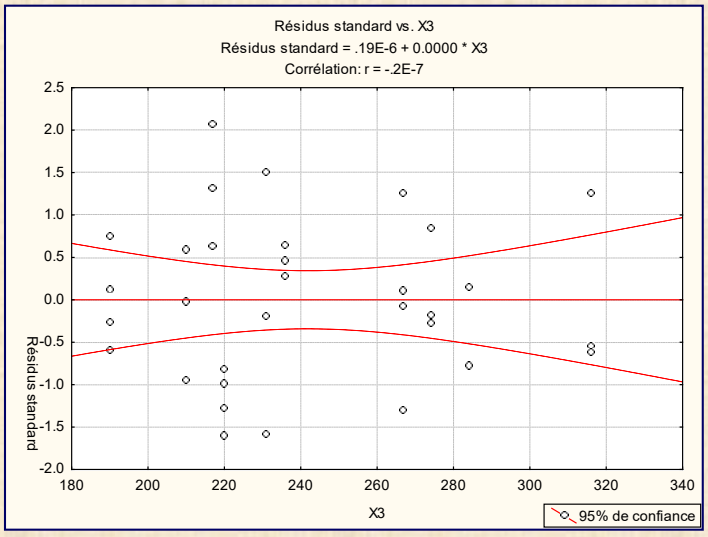
résidu vs X1



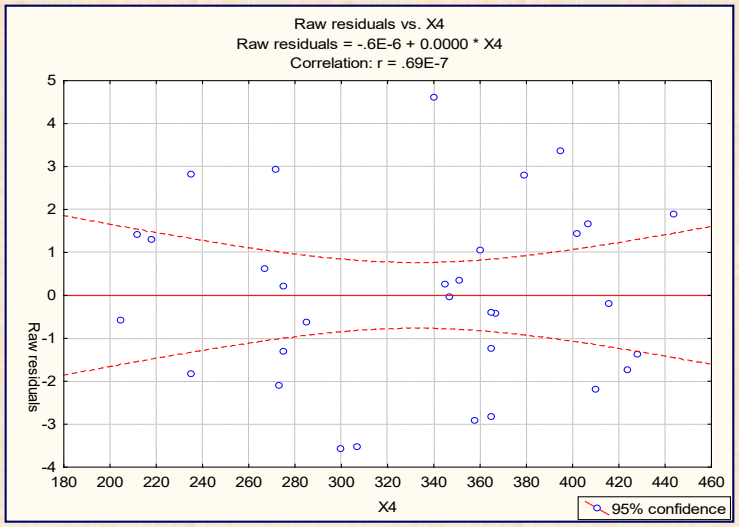
résidu vs X2



résidu vs X3



résidu vs X4



RÉGRESSION LINÉAIRE MULTIPLE (17 / 17)

CORRECTIFS si les résidus présentent des anomalies

- **élimination** variables colinéaires redondantes :
  méthode de sélection de variables pas à pas (stepwise)
- **ajout** termes additionnels dans modèle : X_i^2 , $X_i X_{i'}$ $(i \neq i')$
- **élimination** d'observations influentes
- **ajout** de nouvelles variables explicatives
- **recherche** de nouveaux modèles / formes fonctionnelles

Transformation de Box-Cox de Y pour stabiliser la variance

nouvelle variable Y' transformation de puissance

$$Y' = Y^{\lambda} \qquad -2 < \lambda < 2$$

bonus fréquent : corrige l'absence de normalité de la réponse Y

méthodes pour obtenir λ : graphique (page suivante) ou analytique (ch. 5)

Forme équivalente de la transformation de Box-Cox

$$Y' = \begin{cases} (Y^{\lambda} - 1) / \lambda\, g^{\lambda - 1} & \lambda \neq 0 \\ g \ln (Y) & \lambda = 0 \end{cases}$$

$$g = \exp\left[(1/n)\sum \ln (y)\right] = \Pi\, y^{1/n}$$ moyenne géométrique

Transformation de Box-Cox

$$Y' = Y^{\lambda} \qquad -2 < \lambda < 2$$

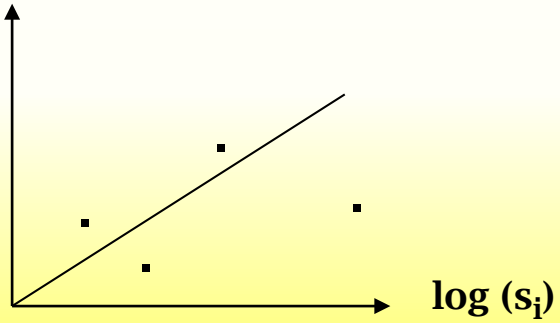
groupes de données $(\bar{Y}_i , s_i)$ $\quad i = 1, 2, 3, \ldots, k$

méthode graphique pour déterminer λ

graphique de $\log(\bar{Y}_i)$ vs $\log(s_i)$ $\quad$ pente = α $\quad \lambda = 1 - \alpha$

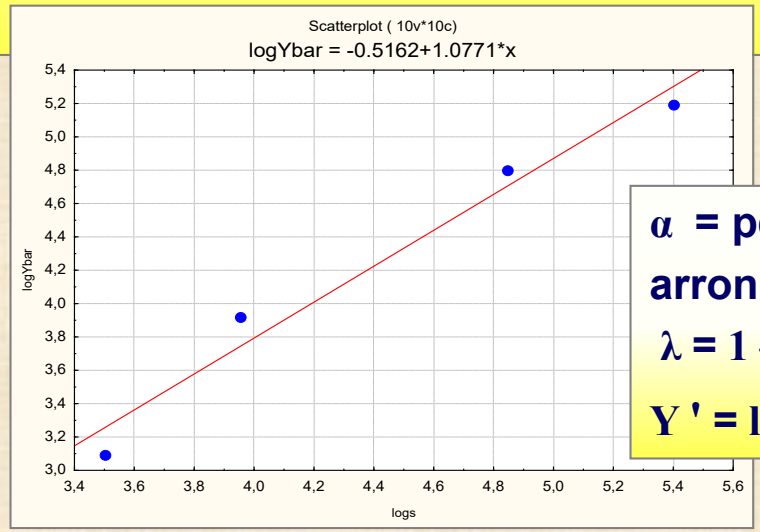
(arrondir)

α	λ	Y'
3	-2	$1 / Y^2$
2	-1	$1 / Y$
1.5	-0.5	$Y^{-0.5}$
1	0	log (Y)
0.5	0.5	$Y^{0.5}$
0	1	Y
-1	2	Y^2



exemple

groupe	Ybar	s
A	22.13	33.22
B	50.37	52.29
C	121.21	127.15
D	180.41	222.14



α = pente = 1.07

arrondi $\alpha = 1$

$\lambda = 1 - \alpha = 0$

$Y' = \log(Y)$

PLAN 2^4 : 16 ESSAIS
4 facteurs variant à 2 modalités

matrice de design					iden.	Réponse
colonne	1	2	3	4		
essai	A	B	C	D	id	Y
1	-	-	-	-	(1)	y1
2	+	-	-	-	a	y2
3	-	+	-	-	b	y3
4	+	+	-	-	ab	y4
5	-	-	+	-	c	y5
6	+	-	+	-	ac	y6
7	-	+	+	-	bc	y7
8	+	+	+	-	abc	y8
9	-	-	-	+	d	y9
10	+	-	-	+	ad	y10
11	-	+	-	+	bd	y11
12	+	+	-	+	abd	y12
13	-	-	+	+	cd	y13
14	+	-	+	+	acd	y14
15	-	+	+	+	bcd	y15
16	+	+	+	+	abcd	y16

- : modalité 1
+ : modalité 2

chaque ligne représente un traitement

identification

présence d'une lettre minuscule implique que le facteur prend la modalité +

ordre standard Yates

colonne 1 (A) alternance des signes - et +

colonne 2 (B) alternance - - + + etc

propriétés importantes du plan factoriel complet 2^4

- 8 (+) et 8 (-) dans chaque colonne
  somme = 0
- ORTHOGONALITÉ ⟶ permet de séparer les effets
  produit de 2 colonnes = 0
- ÉQUILIBRÉ «balance»
  - chaque modalité (niveau) de chaque facteur apparaît exactement 8 fois
  - toutes les combinaisons de 2 facteurs apparaissent exactement 4 fois
  - toutes les combinaisons de 3 facteurs apparaissent exactement 2 fois

PLAN DE 16 ESSAIS 2^4 : matrice des effets – modèle pour l'analyse

		plan				interactions doubles						interactions triples				inter. quadruple	
#	gen	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD	Y
1	1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1	y1
2	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	y2
3	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1	y3
4	1	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	y4
5	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	y5
6	1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	y6
7	1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	y7
8	1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	y8
9	1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1	y9
10	1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	y10
11	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	y11
12	1	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	y12
13	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	y13
14	1	1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	y14
15	1	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1	y15
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	y16

PLAN DE 16 ESSAIS 2^4 : matrice des effets – modèle pour l'analyse

- colonnes AB, AC, BC, ABC,… sont obtenues par multiplication – AB = A x B, … BCD =B x C x D ….
- engendre un maximum de 15 comparaisons appelés CONTRASTES

$$C = \sum c_i Y_i \quad \text{où} \quad \sum c_i = 0 \qquad c_i = \pm 1$$

- les contrastes sont orthogonaux : produit de 2 colonnes = 0

séparation complète des effets principaux et d'interaction

- les contrastes transforment la réponse Y pour obtenir
  - 4 effet principaux : A - B - C - D
  - 6 effets d'interactions doubles : AB - AC - AD – BC – BD -CD
  - 4 effets d'interaction triples : ABC – ABD – ACD - BCD
  - l'interaction quadruple ABCD
- possibilité d'employer la matrice avec

  5 – 6 – 7 – … – 14 – 15 facteurs : plans fractionnaires 2^{k-p}

Exploitation de la matrice augmentée 16 essais

Application 1 : modèle avec 4 facteurs A, B,C, D

notation : $X_{AB} = X_A X_B$, $X_{ABC} = X_A X_B X_C$ etc.

modèle

$$Y = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D$$
$$+ \beta_{AB} X_{AB} + \beta_{AC} X_{AC} + \beta_{AD} X_{AD} + \beta_{BC} X_{BC} + \beta_{BD} X_{BD} + \beta_{CD} X_{CD} +$$
$$+ \beta_{ABC} X_{ABC} + \beta_{ABD} X_{ABD} + \beta_{ACD} X_{ACD} + \beta_{BCD} X_{BCD} +$$
$$+ \beta_{ABCD} X_{ABCD}$$

$$Y = \sum_{j=0}^{15} \beta_j X_j$$

où $X_0 = 1$, $X_1 = X_A$, $X_2 = X_B$, ... , $X_{15} = X_A X_B X_C X_D$

colonne identité X_0 + les 15 colonnes de la matrice

Exploitation de la matrice augmentée 16 essais

Application 2 : modèle avec 5 facteurs A, B,C, D, E en 16 essais

5ième facteur E est définit (confondu) avec l'interaction , E = ABCD

$$Y = \sum_{j=0}^{15} \beta_j X_j$$

où X_j sont les 15 colonnes de la matrice $j = 1, 2, \ldots, 15$

et $X_0 = 1$ la colonne identité

conséquences de E = ABCD

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
A	B	C	D	AB	AC	AD	BC	BD	CD	~~ABC~~	~~ABD~~	~~ACD~~	~~BCD~~	~~ABCD~~
				~~CDE~~	~~BDE~~	~~BCE~~	~~ADE~~	~~ACE~~	~~ABE~~	DE	CE	BE	AE	E

modèle : effets principaux et les interactions doubles seulement toutes séparées -- plan de résolution V

$$Y = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_{AB} X_{AB} + \beta_{AC} X_{AC} + \beta_{AC} X_{AD} + \beta_{BC} X_{BC} + \beta_{BD} X_{BD} + \beta_{CD} X_{CD} + \beta_{DE} X_{DE} + \beta_{CE} X_{CE} + \beta_{BE} X_{BE} + \beta_{AE} X_{AE} + \beta_E X_E$$

Exploitation de la matrice augmentée 16 essais

Application 3 : modèle 8 facteurs A, B,C, D, E, F, G, H en 16 essais

E, F, G, H définies (confondues) avec les 4 interactions triples :

E = BCD F = ACD G = ABC H = ABD (chapitre 3-27)

Modèle $Y = \sum_{j=0}^{15} \beta_j X_j$ X_j 15 colonnes de la matrice j = 1,..,15

$X_0 = 1$ colonne identité

conséquences : définition de E, F, G , H avec des interactions triples

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
A	B	C	D	AB	AC	AD	BC	BD	CD	~~ABC~~	~~ABD~~	~~ACD~~	~~BCD~~	~~ABCD~~
				CG	BG	BH	AG	AH	AF	G	H	F	E	AE
				DH	DF	CF	DE	CE	BE					BF
				EF	EH	EG	FH	FG	GH					CH
														DG

colonnes 5 - 6 -7- 8 – 9 – 10 – 15 : interactions doubles confondues

complique l'interprétation si le coefficient β associé à la colonne est «grand»

aide : identifier effets principaux «importants» - colonnes 1-2- 3- 4-11-12-13-14

$$Y = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_5 (X_{AB} + X_{CG} + X_{DH} + X_{EF})$$
$$+ \beta_6 (X_{AC} + X_{BG} + X_{DF} + X_{EH}) + \ldots + \beta_{10} (X_{CD} + X_{AF} + X_{BE} + X_{GH})$$
$$+ \beta_{11} X_G + \beta_{12} X_H + \beta_{13} X_F + \beta_{14} X_E + \beta_{15} (X_{AE} + X_{BF} + X_{CH} + X_{DG})$$

Analyse du plan de 16 essais

modèle et données $y_i = \sum_{j=0}^{15} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, 2, \ldots, 16$

y_i : réponse observée à l'essai (traitement i)

β_j : coefficients du modèle (à estimer) $\quad j = 0, 1, .., 15$

x_{ij} : constantes connues (± 1) - colonnes de la matrice

$x_{i0} = 1 \qquad i = 1, 2, ,\ldots, 16$

Propriétés matrice X ⟶ orthogonale

$$\sum x_{ij} = 0 \qquad \sum x_{ij}^2 = 16 \qquad \sum x_{ij}\, x_{ij'} = 0 \qquad j \neq j'$$

conséquence : simplification des calculs de régression (ch. 4.6)

n données (répétitions) essai i : $y_{i1}, y_{i2}, \ldots, y_{in}$

$\bar{y}_i = \sum_{k}^{n} y_{ik} / n$: moyenne de l'essai i

$s_i^2 = \sum (y_{ik} - \bar{y}_i)^2 / (n - 1)$: variance de l'essai

$\bar{\bar{y}} = \sum_{k} \bar{y}_i / 16$: grande moyenne

Analyse du plan de 16 essais

Estimation β **moindres carrés** **formules p.6**

$$\hat{\beta}_0 = \bar{y} \qquad \hat{\beta}_j = \sum x_{ij} \bar{y}_i \; / 16$$

Modèle ajusté $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i\,1} + \hat{\beta}_2 x_{i\,2} + \ldots + \hat{\beta}_{15} x_{i\,15}$

Estimation de l'erreur expérimentale σ

$\hat{\sigma}^2 = \sum s_i^2 / 16$ estimation directe (appelée erreur pure)

applicable seulement si n > 1

si n = 1 plan sans répétition

l'estimation de σ est indirecte et basée sur la somme des carrés résiduels (s'il reste des degrés de liberté) après l'ajustement du modèle: dépend du modèle employé

méthode pour juger de l'importance des effets sans faire des tests de signification

effets placés sur «half normal prob plot» : détails chapitre 5

Analyse du plan de 16 essais

ANOVA

$SS_{TOT} = \sum \sum (y_{ik} - \bar{y})^2$: variabilité totale

--- $SS_{MOD} = n \sum (\hat{y}_i - \bar{y})^2$: variabilité expliquée modèle (inter)

$SS_{INTRA} = \sum \sum (y_{ik} - \bar{y}_i)^2$: intra variabilité $= 16\,\hat{\sigma}^2$

$SS_{TOT} = SS_{MOD} + SS_{INTRA}$: décomposition de la variabilité

--> $SS_{MOD} = 16n \sum \hat{\beta}_j^2$: décomposition orthogonale des effets

degrés de liberté

$df_{TOT} = 16n - 1$: degrés de liberté de SS_{TOT}

$df_{MOD} = 15$: degrés de liberté de SS_{MOD}

$df_{INTRA} = 16(n - 1)$: degrés de liberté de SS_{INTRA}

Analyse du plan de 16 essais

Tableau d'analyse de la variance

source	somme carrés (SS)	degrés liberté (df)	carrés moyens (MS)	F (ou t) (*)
modèle	SS_{MOD}	15	MS_{MOD}	$MS_{MOD} / \hat{\sigma}^2$
x_1	$16n\,\hat{\beta}_1^2$	1	$16n\,\hat{\beta}_1^2$	$16n\,\hat{\beta}_1^2 / \hat{\sigma}^2$
.	.	1	.	.
x_{15}	$16n\,\hat{\beta}_{15}^2$	1	$16n\,\hat{\beta}_{15}^2$	$16n\,\hat{\beta}_{15}^2 / \hat{\sigma}^2$
intra	SS_{INTRA}	16 (n-1)	$MS_{INTRA} = \hat{\sigma}^2$	----------
totale	SS_{TOT}	16n – 1	----------	

(*) Les tests de signification F des coefficients β de chaque x sont équivalents aux tests t de Student

Analyse du plan de 16 essais

Test

global $H_0 : \beta_1 = \beta_2 = \ldots\ldots = \beta_{15} = 0$ aucun facteur est actif sur Y

rejet de H_0 si $MS_{MOD} / \hat{\sigma}^2 > F_{15\ ,\ 16\,(n-1)\ ,\ 1-\alpha}$ Fisher

individuel $H_{0j} : \beta_j = 0$ effet principal / interaction est zéro

rejet de H_{0j} si $16\, n\, \hat{\beta}_1^2 / \hat{\sigma}^2 > F_{1\ ,\ 16\,(n-1)\ ,\ 1-\alpha} = t^2_{16(n-1),\ 1-\alpha}$

Variation expliquée

$$R^2 = SS_{MOD} / SS_{TOT} \qquad 0 \leq R^2 \leq 1$$

$$R_{adj}^2 = 1 - [\,(16n - 1) / (16n - 15)\,]\ (\,1 - R^2\,)$$

Analyse du plan de 16 essais

décomposition de la variabilité

$SS_{TOT} = SS_{INTER} + SS_{INTRA}$ 1 facteur (chapitre 2)

$SS_{TOT} = SS_{MOD} + SS_{RESID}$ régression (chapitre 4)

$SS_{TOT} = SS_{MOD} + SS_{INTRA}$ plan 16 essais / 15 effets / n > 1

Cas général (si n > 1 : répétitions)

$$SS_{TOT} = SS_{MOD} + SS_{RESID}$$

$$= SS_{MOD} + SS_{INTRA} + SS_{LOF}$$

LOF : « lack of fit »

exemples : chapitres 5/6

différence effet significatif VERSUS effet important

RELATION entre F et R^2

$$R^2 = df\_modèle * F\_modèle / (df\_modèle * F\_modèle + df\_erreur)$$

$$F = df\_erreur * R^2 / (df\_modèle * (1 - R^2))$$

Un ratio F_modèle dépassant le F_critique à 5% (F_critique) permet de rejeter l'hypothèse nulle il n'implique pas nécessairement un R^2 élevé comme on peut le constater dans le tableau.
Il faut distinguer entre un effet (test) SIGNIFICATIF et un effet qui est IMPORTANT disons un R^2 supérieur à 50%. Il faut que le ratio F soit au moins 4 fois plus grand (F_gros = 4*F_critique (0,05) pour que le R^2 dépasse 50% .

df model	df erreur	F Critique (0,05)	R^2	F_gros = 4 * F critique	R^2 F gros
1	5	6,61	0,57	26,43	0,84
1	10	4,96	0,33	19,86	0,67
1	15	4,54	0,23	18,17	0,55
1	20	4,35	0,18	17,40	0,47
1	30	4,17	0,12	16,68	0,36
1	50	4,03	0,07	16,14	0,24
3	5	5,41	0,76	21,64	0,93
3	10	3,71	0,53	14,83	0,82
3	15	3,29	0,40	13,15	0,72
3	20	3,10	0,32	12,39	0,65
3	30	2,92	0,23	11,69	0,54
3	50	2,79	0,14	11,16	0,40
5	5	5,05	0,83	20,20	0,95
5	10	3,33	0,62	13,30	0,87
5	15	2,90	0,49	11,61	0,79
5	20	2,71	0,40	10,84	0,73
5	30	2,53	0,30	10,13	0,63
5	50	2,40	0,19	9,60	0,49
10	5	4,74	0,90	18,94	0,97
10	10	2,98	0,75	11,91	0,92
10	15	2,54	0,63	10,17	0,87
10	20	2,35	0,54	9,39	0,82
10	30	2,16	0,42	8,66	0,74
10	50	2,03	0,29	8,10	0,62