MODÈLE LINÉAIRE

Plan du Chapitre 1

1. Les données
2. Présentation d'exemples
3. Écriture d'un modèle
4. Cadre du modèle linéaire
5. Postulats
6. Liens avec les autres modules de FPSTAT2

Les données

On dispose d'un ensemble de N observations sur lesquelles ont été effectuées V + 1 mesures des variables Y, Z1, Z2, ..., ZV.

On veut expliquer ou prévoir Y à l'aide des variables Z1, Z2, ..., ZV.

Y est une variable **quantitative** appelée *variable réponse* ou *variable dépendante* ou *variable expliquée* ou encore *variable endogène*.

Z1, Z2, ..., ZV sont des variables **qualitatives** ou **quantitatives** appelées *prédicteurs* ou *variables indépendantes* ou *variables explicatives* ou encore *variables exogènes*.

Les *prédicteurs* **quantitatifs** sont aussi appelés **régresseurs**.

Les *prédicteurs* **qualitatifs** sont aussi appelés **facteurs**.

Présentation d'exemples

Exemple 1 : Étude des forêts à travers la hauteur d'arbres.

Hauteur	Forêt
23.4	1
24.4	1
24.6	1
24.9	1
25.0	1
26.2	1
18.9	2
21.1	2
21.1	2
22.1	2
22.5	2
23.5	2
24.5	2
22.5	3
22.9	3
23.7	3
24.0	3
24.0	3

Question : Nature des variables ? Méthode d'analyse ?

Exemple 2 : Étude de la teneur en huile de populations de tournesol d'origines variées, croisées à deux testeurs T1 et T2

Teneur en huile	Testeur	origine
43.54	T1	Afrique
45.30	T1	Afrique
44.25	T1	Hongrie
42.55	T1	Hongrie
47.28	T1	Maroc
49.40	T1	Maroc
47.21	T2	Afrique
47.73	T2	Afrique
44.34	T2	Hongrie
46.49	T2	Hongrie
47.75	T2	Maroc
49.47	T2	Maroc

Question : Nature des variables ? Méthode d'analyse ?

Age (Z)	Tension (Y)
35	114
45	124
55	143
65	158
75	166

Question : Nature des variables ? Méthode d'analyse ?

Exemple 4 : Étude du rendement de blé en fonction de doses de fertilisants AZ, PH et PO

RENDEMENT	AZ	PH	PO
30	80	40	40
50	100	40	40
100	180	100	100
60	100	80	20
70	150	70	120

Question : Nature des variables ? Méthode d'analyse ?

variété de moutarde à la sécheresse

Mesures 34 jours après repiquage (sans irrigation).

RC : Nombre de racines courtes tubérisées.

LT : Longueur de la tige.

HF : Potentiel hydrique foliaire.

PR : Poids matière sèche des racines.

PA : Poids matière sèche des parties aériennes.

FE : Nombre de feuilles.

OBS	RC	LT	HF	PR	PA	FE
1	0.00	29	65	87	43	2
2	0.00	35	65	163	122	2
3	1.10	40	65	175	117	3
4	0.69	25	60	38	49	2
5	0.00	30	30	57	23	1
6	0.00	45	70	270	124	5
7	0.69	40	65	202	78	4
8	1.39	50	70	226	74	3
9	1.61	50	85	525	222	5
10	1.10	55	80	230	92	3
11	3.47	60	155	1109	897	4
12	2.40	80	95	869	628	5
13	1.10	60	60	553	189	8
14	3.00	90	100	903	3022	6
15	3.43	80	145	1216	3049	6
16	1.61	75	85	912	3273	6
17	2.83	60	75	689	443	6
18	1.61	85	85	443	251	5
19	2.20	65	80	643	424	5
20	4.09	60	240	1089	843	6
21	3.09	60	80	825	757	7
22	1.39	70	80	385	1350	5
23	4.22	90	180	1335	728	7
24	4.17	90	175	953	668	3
25	4.57	95	205	1145	696	6
26	3.43	75	305	1129	678	6
27	3.04	70	120	978	529	6
28	3.26	75	70	795	329	6
29	5.40	70	300	1618	1075	7
30	4.16	70	250	1020	881	7
31	3.91	60	280	1020	624	8

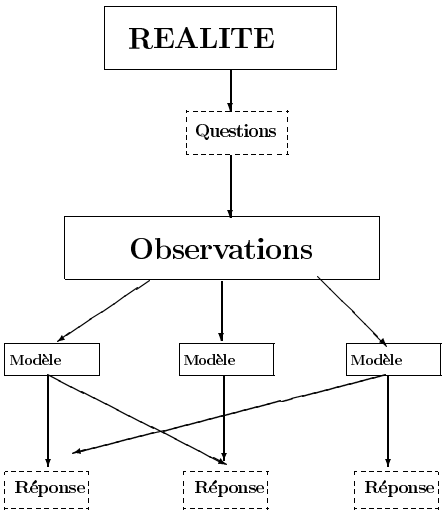
Question : Nature des variables ? Méthode d'analyse ?

en fonction du traitement (dose) d'un fongicide

OBS	TRAITEMENT	DOSE(kg/ha)	RENDEMENT(g/pied)
1	T1	0.25	377
2	T1	0.25	408
3	T1	0.25	500
4	T1	0.25	333
5	T2	0.50	527
6	T2	0.50	604
7	T2	0.50	606
8	T2	0.50	533
9	T3	0.75	623
10	T3	0.75	550
11	T3	0.75	562
12	T3	0.75	667
13	T4	1.00	633
14	T4	1.00	600
15	T4	1.00	650
16	T4	1.00	567
17	T5	1.25	642
18	T5	1.25	708
19	T5	1.25	662
20	T5	1.25	504

Question : Nature des variables ? Méthode d'analyse ?

Pourquoi un modèle ?



Exemple 1: Étude des forêts à travers la hauteur d'arbres

facteur forêt : i (1 à I = 3)

répétition : r (1 à n_i)

Modèle

$$Y_{ir} = \mu + \alpha_i + \epsilon_{ir}$$

$$\begin{cases} Y_{1r} = & 1.\theta_0 + 1.\theta_1 + 0.\theta_2 + 0.\theta_3 + \epsilon_{1r} & r = 1, \cdots, 6 \\ Y_{2r} = & 1.\theta_0 + 0.\theta_1 + 1.\theta_2 + 0.\theta_3 + \epsilon_{2r} & r = 1, \cdots, 7 \\ Y_{3r} = & 1.\theta_0 + 0.\theta_1 + 0.\theta_2 + 1.\theta_3 + \epsilon_{3r} & r = 1, \cdots, 5 \\ & \downarrow \quad \downarrow \quad \downarrow \quad \downarrow & \\ & x_{0ir} \quad x_{1ir} \quad x_{2ir} \quad x_{3ir} & \end{cases}$$

Exemple 2 : Étude de la teneur en huile de populations de tournesol

facteur testeur : i (1 à I = 2) facteur origine : j (1 à J = 3)

répétition : r (1 à $n_{ij} = 2$)

Modèle : sans interaction

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \epsilon_{ijr}$$

$$\begin{cases} Y_{11r} = & 1.\theta_0 + 1.\theta_1 + 0.\theta_2 + 1.\theta_3 + 0.\theta_4 + 0.\theta_5 + \epsilon_{11r} \\ Y_{12r} = & 1.\theta_0 + 1.\theta_1 + 0.\theta_2 + 0.\theta_3 + 1.\theta_4 + 0.\theta_5 + \epsilon_{12r} \\ Y_{13r} = & 1.\theta_0 + 1.\theta_1 + 0.\theta_2 + 0.\theta_3 + 0.\theta_4 + 1.\theta_5 + \epsilon_{13r} \\ Y_{21r} = & 1.\theta_0 + 0.\theta_1 + 1.\theta_2 + 1.\theta_3 + 0.\theta_4 + 0.\theta_5 + \epsilon_{21r} \\ Y_{22r} = & 1.\theta_0 + 0.\theta_1 + 1.\theta_2 + 0.\theta_3 + 1.\theta_4 + 0.\theta_5 + \epsilon_{22r} \\ Y_{23r} = & 1.\theta_0 + 0.\theta_1 + 1.\theta_2 + 0.\theta_3 + 0.\theta_4 + 1.\theta_5 + \epsilon_{23r} \\ & \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow & \\ & x_{0ijr} \quad x_{1ijr} \quad x_{2ijr} \quad x_{3ijr} \quad x_{4ijr} \quad x_{5ijr} & \end{cases}$$

huile de populations de tournesol

facteur testeur : i (1 à I = 2) facteur origine : j (1 à J = 3)

répétition : r (1 à n_{ij} = 2)

Modèle : avec interaction

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijr}$$

$$\begin{cases} Y_{11r} = & 1.\theta_0 + 1.\theta_1 + 0.\theta_2 + 1.\theta_3 + 0.\theta_4 + 0.\theta_5 + \\ & 1.\theta_6 + 0.\theta_7 + 0.\theta_8 + 0.\theta_9 + 0.\theta_{10} + 0.\theta_{11} \quad + \epsilon_{11r} \\ \vdots & \\ Y_{23r} = & 1.\theta_0 + 0.\theta_1 + 1.\theta_2 + 0.\theta_3 + 0.\theta_4 + 1.\theta_5 + \\ & 0.\theta_6 + 0.\theta_7 + 0.\theta_8 + 0.\theta_9 + 0.\theta_{10} + 1.\theta_{11} \quad + \epsilon_{23r} \end{cases}$$

Exemple 3 : Étude de la relation entre la tension artérielle et l'âge

$$Y_n = \alpha + \beta.Z_n + \varepsilon_n \quad n = 1 \cdots N$$
$$Y_n = 1.\theta_0 + Z_n\theta_1 + \varepsilon_n \quad n = 1 \cdots N$$

Exemple 4 : Étude du rendement de blé en fonction de doses de fertilisants AZ, PH et PO

$$Y_n = \beta_0 + \beta 1.Z1_n + \beta 2.Z2_n + \beta 3.Z3_n + \epsilon_n$$
$$Y_n = 1.\theta_0 + Z1_n\theta_1 + Z2_n\theta_2 + Z3_n\theta_3 + \epsilon_n$$

Exemple 5 : Étude de l'adaptation d'une variété de moutarde à la sécheresse

$$Y_n = 1.\theta_0 + Z1_n\theta_1 + Z2_n\theta_2 + Z3_n\theta_3 + Z4_n\theta_4 + Z5_n\theta_5 + \epsilon_n$$

Exemple 6 : Étude du rendement de pomme de terre en fonction du traitement (dose) d'un fongicide

→ 1er modèle : analyse de la variance

facteur traitement (dose) : i (1 à I = 5)

répétition : r (1 à n_i = 4)

$$Y_{ir} = \mu + \alpha_i + \epsilon_{ir}$$

$$\begin{cases} Y_{1r} = \mathbf{1}.\theta_0 + \mathbf{1}.\theta_1 + \mathbf{0}.\theta_2 + \mathbf{0}.\theta_3 + \mathbf{0}.\theta_4 + \mathbf{0}.\theta_5 + \epsilon_{1r} \\ \vdots \qquad\qquad\qquad \vdots \\ Y_{5r} = \mathbf{1}.\theta_0 + \mathbf{0}.\theta_1 + \mathbf{0}.\theta_2 + \mathbf{0}.\theta_3 + \mathbf{0}.\theta_4 + \mathbf{1}.\theta_5 + \epsilon_{5r} \end{cases}$$

Exemple 6 : Étude du rendement de pomme de terre en fonction du traitement (dose) d'un fongicide

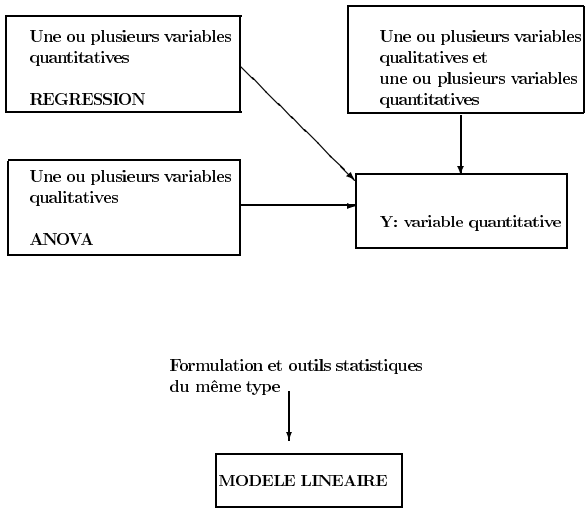
→ 2ème modèle : décomposer le facteur traitement (dose) en une régression linéaire simple

$$Y_{ir} = \mu + (\beta.Z_i + \alpha'_i) + \epsilon_{ir}$$

$$\begin{cases} Y_{1r} = \mathbf{1}.\theta_0 + Z_1.\theta_1 + \mathbf{1}.\theta_2 + \mathbf{0}.\theta_3 + \mathbf{0}.\theta_4 + \mathbf{0}.\theta_5 + \mathbf{0}.\theta_6 + \epsilon_{1r} \\ \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ Y_{5r} = \mathbf{1}.\theta_0 + Z_5.\theta_1 + \mathbf{0}.\theta_2 + \mathbf{0}.\theta_3 + \mathbf{0}.\theta_4 + \mathbf{0}.\theta_5 + \mathbf{1}.\theta_6 + \epsilon_{5r} \end{cases}$$

→ 3ème modèle : décomposer le facteur traitement (dose) en une régression linéaire quadratique

$$Y_{ir} = \mu + \left(\beta_2.Z_i^2 + \beta_1.Z_i + \alpha''_i\right) + \epsilon_{ir}$$



Synthèse sur l'écriture d'un modèle linéaire

$$Y_n = \sum x_{pn}.\theta_p + \epsilon_n$$

$$\mu_n = \sum x_{pn}.\theta_p$$

— paramètres θ_p inconnus à estimer

— μ_n linéaire en les paramètres

Cadre du modèle linéaire

Dans tous les exemples précédents, on a choisi d'approcher Y_n par un modèle statistique particulier : le *Modèle linéaire*.

Choisir le *Modèle linéaire* c'est définir un cadre général de travail en

— optant pour un modèle linéaire sur l'espérance des Y_n

— faisant des postulats sur la loi des ϵ_n

Qu'est ce qu'un modèle linéaire sur l'espérance ?

Définir une équation mathématique sur μ_n

$$\mu_n = \sum x_{pn} \,.\, \theta_p$$

Exemples d'équation de modèles linéaires :

$\mu_n = \theta_0 + Z1_n.\theta_1 + (Z1_n)^2.\theta_2$
$\mu_n = \theta_0 + \log(Z1_n).\theta_1 + e^{Z2_n}.\theta_2$

Contre-exemple d'équation de modèles linéaires :

$$\mu_n = \theta_0 + \frac{1}{Z1_n + \theta_1}.\theta_2$$

Parmi ces modèles, quels sont ceux qui sont linéaires ?

1. $\mu_n = \theta_0 + \sin(Z1_n) \,.\, \theta_1 + (Z1_n/Z2_n) \,.\, \theta_2$
2. $\mu_n = \sqrt{Z1_n} \,.\, \theta_0$
3. $\mu_n = \theta_0 \,.\, e^{-\theta_1}$
4. $\cdots$

Postulats du modèle linéaire

Définitions

Les différents postulats du *Modèle Linéaire* sont supposés **vrais au départ.**

Ils concernent la loi conjointe des ϵ_n.

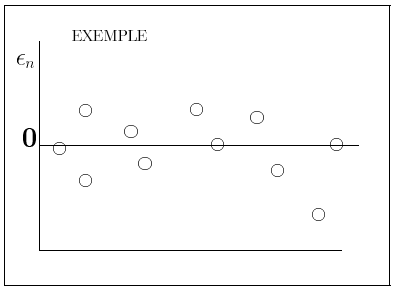
— **P1 :** ils ont une espérance nulle : $E(\epsilon_n) = 0$;

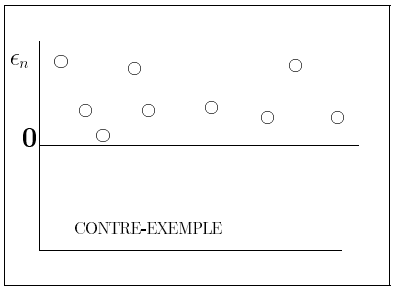
— **P2** : ils ont tous même variance : $\mathrm{Var}(\epsilon_n) = \sigma^2$ (Homoscédasticité) ;

— **P3** : ils sont deux à deux non corrélés : $\mathrm{Cov}(\epsilon_n, \epsilon_{n'}) = 0 \,\mathrm{pour}\, n \neq n'$ (indépendance) ;

— **P4 :** ils suivent des lois gaussiennes : $\mathcal{N}(0, \sigma^2)$

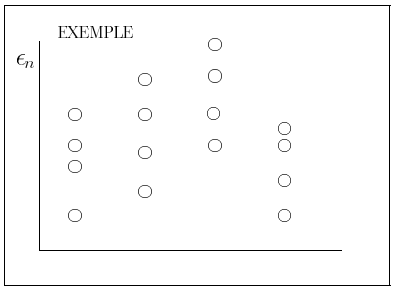
$E(\epsilon_n) = 0$

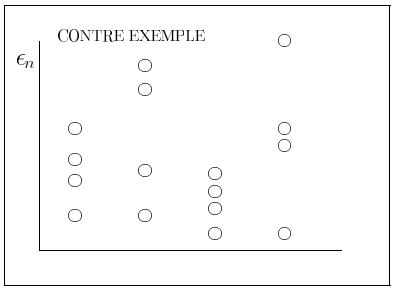




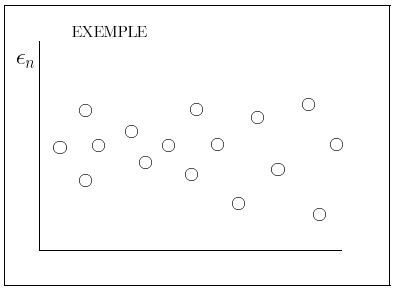
Postulats du modèle linéaire

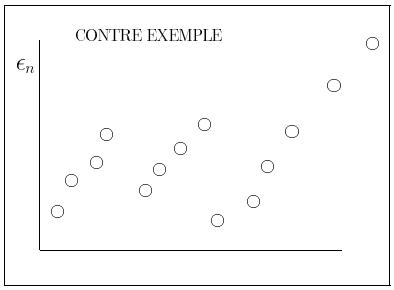
$\mathbf{Var}(\epsilon_n) = \sigma^2$



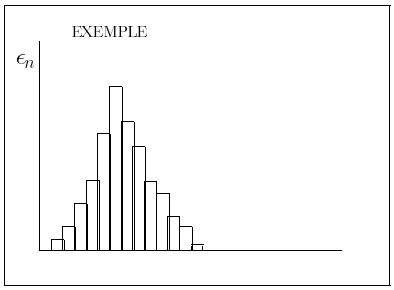


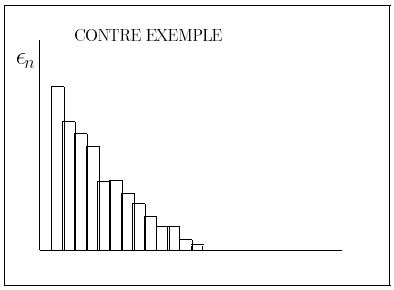
$\mathbf{Cov(\epsilon_n, \epsilon_{n'}) = 0 \text{ pour } n \neq n'}$





Postulats du modèle linéaire
$\mathcal{N}(0, \sigma^2)$





Choisir le cadre du Modèle linéaire, c'est se donner

— une équation mathématique sur la forme de $\mu_n = \sum x_{pn} \, . \, \theta_p$

— des postulats sur les ϵ_n

Si μ_n n'est pas linéaire dans ses paramètres, il faut envisager la possibilité d'un modèle non-linéaire.

Si ce n'est pas μ_n, mais $g(\mu_n)$ qui est linéaire dans ses paramètres et si la loi des ϵ_n est par exemple binomiale ou poissonnière, on entre dans le cadre du Modèle Linéaire Généralisé (GLM).

Si la variance n'est pas constante, on peut transformer les variables pour stabiliser la variance.

Si la variance et la covariance ne sont pas constantes et sous certaines conditions, on entre dans le cadre du Modèle Mixte.