INTRODUCTION AU 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 objectifs liés aux différents points sont les suivants :

— **Données** définir le cadre général dans lequel on se place en particulier pour tout ce qui concerne la nature des variables, leurs rôles, ...

— **Exemples** présenter des exemples réels, supposés connus des stagiaires et sur lesquels on va travailler tout au long du stage.

— **Écriture** conclure sur l'existence d'un modèle mathématique unique permettant de traiter tous les exemples vus précédemment. Donner son écriture et le cadre général de son utilisation.

— **Modèle Linéaire sur l'Espérance** définition d'un ML

— **Postulats** Explicitation des Postulats que l'on fait quand on utilise le ML

— **Liens** alternatifs ou aménagement du ML

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-1

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**.

Les données : définition du cadre général dans lequel on se place.

On dispose d'un ensemble de données sur lesquelles ont été mesurées un certain nombre de variables. Les variables mesurées sont de nature différente :

— une variable **quantitative** Y, objet de l'étude, que l'on cherche à expliquer, prédire, ... Cette variable est appelée *variable réponse* , mais elle a aussi d'autre noms selon les disciplines : *dépendante, variable expliquée, variable endogène* , ...

— une ou des variables Z1, Z2, ..., ZV appelée(s) *prédicteur(s)*. Ces variables sont soit **qualitatives** et on parlera de **facteurs**, soit **quantitatives** et on parlera de **régresseurs**. Ces variables sont supposées connues sans erreur.

— Les deux types de variables ne jouent pas *a priori* le même rôle : les valeurs prises par la *variable réponse* Y ne sont pas connues et dépendent des valeurs prises par les Z1, Z2, ..., ZV *prédicteurs*.

Remarque : On ne s'intéressera au cours du stage qu'au cas d'une seule variable réponse. L'application du modèle linéaire au cas multivariable n'est pas abordée.

Remarque : Dans la pratique, il est souvent difficile de fixer les valeurs des *prédicteurs*, en particulier dans le cas des **régresseurs** (les variables Y et Z1, Z2, ..., ZV étant bien souvent échantillonnées simultanément).

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 ?

Dans cet exemple, il s'agit de savoir si la moyenne des hauteurs d'arbres est significativement différente dans les trois types de forêt.

— La variable HAUTEUR est continue et **quantitative**, c'est une *variable réponse ou variable expliquée*.

— La variable FORÊT est **qualitative** (type 1, 2 ou 3), c'est une *variable prédicteur, ou variable explicative*.

Les variables **quantitatives** concernent les données numériques telles que les mesures (poids, hauteurs, ...) et la numération (nombre de feuilles, ...). Par contre, les variables **qualitatives** concernent les attributs comme le sexe, la race, la couleur, ...

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 ?

On étudie l'importance de l'origine des tournesols et des testeurs sur la teneur en huile des tournesols. On a deux facteurs de variation :

— Le testeur (lignée = variété homozygote ou hybride F_1 (issu du croisement de 2 lignées)) avec lequel les variétés ont été croisées (T1 et T2) ;

— l'origine géographique des variétés (Afrique, Hongrie, Maroc) .

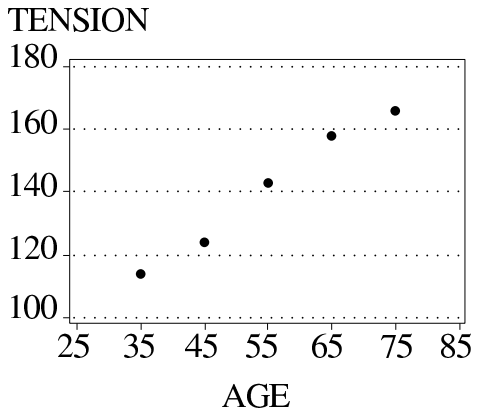
Les variables TESTEUR et ORIGINE sont **qualitatives,** et ce sont des *variables prédicteurs ou variables explicatives*. La variable TENEUR EN HUILE est continue et **quantitative,** c'est une *variable réponse, ou variable expliquée*.

Remarque :

Un testeur est utilisé pour étudier le comportement en croisement de variétés (ou d'autres génotypes)

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

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



Question : Nature des variables ? Méthode d'analyse ?

On s'intéresse à la relation entre la tension artérielle et l'âge. Les variables ÂGE et TENSION sont continues et **quantitatives.** La tension est expliquée par l'âge, ainsi la *variable expliquée* est la TENSION et la *variable explicative* est l'ÂGE.

Remarque : Le graphique montre la relation linéaire entre les variables.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-5

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 ?

On cherche à expliquer ou à prédire le rendement de blé en fonction de doses de fertilisant**.** Les variables RENDEMENT *(variable expliquée)***,** AZ *(variable explicative)*, PH *(variable explicative)*, PO *(variable explicative)* sont **quantitatives**.

Exemple 5 : Étude de l'adaptation d'une 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.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-7

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 ?

Pour étudier l'adaptation d'une variété de moutarde à la sécheresse, on s'intéresse au nombre de racines courtes tubérisées. Une racine tubérisée est formée d'un ou plusieurs tubercules, renflement de la racine. On cherche à expliquer le nombre moyen de racines courtes tubérisées (RC) en fonction de :

— la longueur de la tige (LT) ;

— le potentiel hydrique foliaire (HF) ;

— le poids de matière sèche des racines (PR) ;

— le poids de matière sèche des parties aériennes (PA) ;

— le nombre de feuilles (FE).

Toutes les variables sont **quantitatives**. La *variable expliquée* est le nombre de racines courtes tubérisées (RC) et les *variables explicatives* sont LT, HF, PR, PA, FE.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-8

Exemple 6 : Étude du rendement de pomme de terre 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 ?

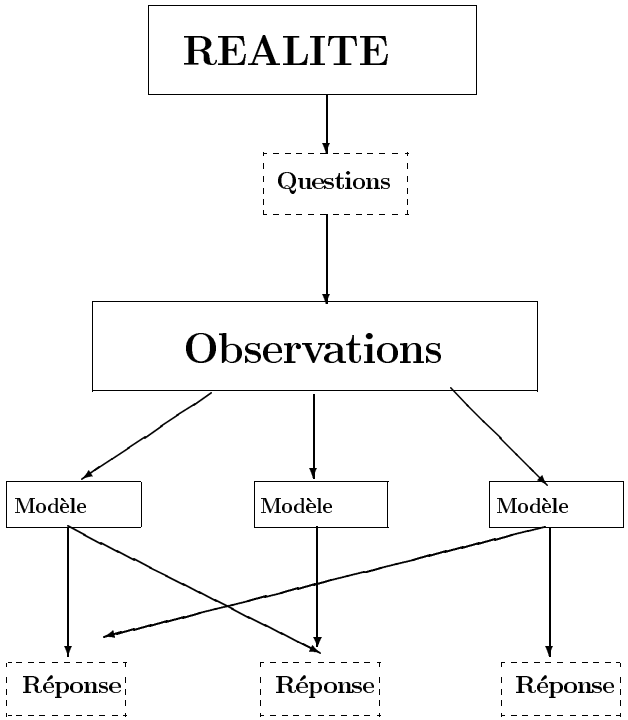
La *variable expliquée* est le rendement (variable **quantitative**) et la *variable explicative* est la dose (variable **qualitative** ou **quantitative**).

Remarque : Dans certains cas, la variable peut être considérée comme **qualitative** s'il s'agit du traitement (T1, T2, T3, T4, T5) ou **quantitative** s'il s'agit de la valeur de la dose (0.25, 0.50, 0.75, 1.00, 1.25).

Les niveaux du traitement (T1, T2, T3, T4, T5) peuvent être quantifiés. Ici, il serait dommage pour l'étude de ne pas prendre en considération le caractère quantitatif des doses et notamment leur croissance.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-9

Pourquoi un modèle ?



L'objectif de base d'un modèle est de construire une représentation simplifiée mais fidèle d'une réalité complexe, permettant de répondre à certaines questions. Parmi les différents types de modèle possibles (mécanistes, logiques, statistiques), nous ne verrons que le modèle statistique.
L'un des modèles statistiques le plus largement utilisé exprime la variable réponse comme une somme de 2 composantes :

variable réponse = composante systématique + composante résiduelle

La *composante systématique* résume la façon dont la variabilité de la variable réponse est prise en compte par les prédicteurs et la *composante résiduelle* résume la part de variabilité non décrite par les prédicteurs.

Le modèle statistique, basé sur des données expérimentales ou observées, est essentiellement empirique. Il va chercher à expliquer les relations existantes entre une variable réponse et une série de prédicteurs et fournir une indication de l'incertitude de sa représentation.

Écriture d'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

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

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

Remarque : Sous la forme de TP, on s'appuiera sur la connaissance pratique de la régression et de l'analyse de la variance pour donner des écritures possibles de modèles. On mettra en évidence la linéarité en les paramètres à l'aide de l'écriture avec θ.
La variable FORÊT est un *prédicteur* **qualitatif**, c'est donc un **facteur** (on note i le niveau du facteur forêt variant de 1 à 3). La répétition r représente l'individu r dans la forêt i, n_i le nombre d'individus dans la forêt i, et N le nombre total d'individus.

Les formulations suivantes sont équivalentes.

— formulation courante : $Y_{ir} = \mu + \alpha_i + \epsilon_{ir}$

— formulation unificatrice : $Y_{ir} = \sum_{p=0}^{C-1} x_{pn}\,\theta_p + \epsilon_{ir}$

(C représente le nombre de paramètres du modèle (C= 4))

Construction du modèle : analyse de variance à un facteur

— μ : moyenne générale
— α_i : effet du niveau i du facteur A : forêt
— ϵ_{ir} : erreur résiduelle (ce qui n'est pas expliqué par le modèle).

Les $x_{pir}(x_{0ir}, x_{1ir}, x_{2ir}, x_{3ir})$ sont les coefficients (déterminés par l'expérience) des paramètres respectifs θ_0, ..., θ_3 à estimer.
Le paramètre θ_0 représente la moyenne μ.
Le paramètre θ_p représente l'effet de la forêt p pour p variant de 1 à 3. Les x_{pir} prennent les valeurs 1 ou 0 selon que le niveau i correspond ou non au paramètre θ_p et x_{0ir} = 1.

$$Y_{1r} = x_{01r}\,.\,\theta_0 + x_{11r}\,.\,\theta_1 + x_{21r}\,.\,\theta_2 + x_{31r}\,.\,\theta_3 + \epsilon_{1r}$$

Tableau des correspondances entre les niveaux i facteur A et les paramètres θ_p :

Niveau i du facteur A	θ_1	θ_2	θ_3
1	1	0	0
2	0	1	0
3	0	0	1

$x_{11r} = 1$ (le niveau i = 1 du facteur forêt correspond au paramètre θ_1)

$$x_{21r} = x_{31r} = 0$$

$$Y_{ir} = \sum_{p=0}^{C-1} x_{pn}\,\theta_p + \epsilon_{ir}$$

Dans le cadre d'une analyse de variance à un facteur, on a :

— n = ir (individu n correspond à l'individu r pour le niveau i)

— paramètre θ_{p} (0 à C-1 = 3) à estimer

— $x_{pn} = \delta_i^p = \begin{cases} 1 \text{ si p correspond au niveau i} \\ 0 \text{ si p ne correspond pas au niveau i} \end{cases}$ (symbole de Kronecker)

et $x_{0n} = 1$

Remarque : Les inconnus à estimer sont les paramètres θ_p, c'est pourquoi le modèle s'écrit plutôt $Y_{1r} = x_{01r}.\theta_0 + \ldots + \epsilon_{1r}$ au lieu de $Y_{1r} = \theta_0\,.\,x_{01r} + \ldots + \epsilon_{1r}$.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-11

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

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

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

Remarque : Sous la forme de TP, on s'appuiera sur la connaissance pratique de la régression et de l'analyse de la variance pour donner des écritures possibles de modèles. On mettra en évidence la linéarité en les paramètres à l'aide de l'écriture avec θ.

Les variables TESTEUR et ORIGINE sont **qualitatives** et ce sont donc des **facteurs**, que l'on indicera par i (pour le facteur testeur : T1 et T2) et par j (pour le facteur origine : Afrique, Hongrie, Maroc). La répétition r représente l'individu r de testeur i et d'origine j , n_{ij} le nombre d'individus de testeur i et d'origine j, et N le nombre total d'individus.

Les formulations suivantes sont équivalentes.

— formulation courante : $Y_{ijr} = \mu + \alpha_i + \beta_j + \epsilon_{ijr}$

— formulation unificatrice : $Y_{ijr} = \sum_{p=0}^{C-1} x_{pn}\,\theta_p + \epsilon_{ijr}$

(C représente le nombre de paramètres du modèle (C = 6))

Construction du modèle : analyse de variance à deux facteurs

— μ : moyenne générale
— α_i : effet du niveau i du facteur A : testeur
— β_j : effet du niveau j du facteur B : origine
— ϵ_{ijr} : erreur résiduelle (ce qui n'est pas expliqué par le modèle).

Les x_{pijr} (x_{0ijr} à x_{5ijr}) sont les coefficients (déterminés par l'expérience) des paramètres respectifs θ_0 à θ_5 à estimer.
Le paramètre θ_0 représente la moyenne μ.
Le paramètre θ_p représente l'effet du testeur p pour p variant de 1 à 2 et l'effet de l'origine pour p variant de 3 à 5. Les x_{pijr} prennent les valeurs 1 ou 0 selon que les niveaux i et j des facteurs A et B correspondent ou non au paramètre θ_p.

Tableau des correspondances entre les niveaux i du facteur A et les paramètres θ_p :

Niveau i du facteur A	θ_1	θ_2
1	1	0
2	0	1

Tableau des correspondances entre les niveaux j du facteur B et les paramètres θ_p :

Niveau j du facteur B	θ_3	θ_4	θ_5
1	1	0	0
2	0	1	0
3	0	0	1

$x_{111r} = x_{311r} = 1$

i=1 pour le niveau du facteur A correspond au paramètre θ_1

j=1 pour le niveau du facteur B correspond au paramètre θ_3

$$Y_{ijr} = \sum_{p=0}^{C-1} x_{pn}\theta_p + \epsilon_{ijr}$$

Dans le cadre d'une analyse de variance à deux facteurs, on a :

— n=ijr (individu n correspond à l'individu r pour les niveaux i et j des facteurs)

— paramètre θ_p (0 à C-1 = 5) à estimer

— $x_{pn} = \delta_{ij}^{p} = \begin{cases} 1 \text{ si p correspond aux niveaux i et j} \\ 0 \text{ si p ne correspond pas aux niveaux i ou j} \end{cases}$ (symbole de Kronecker)

— $x_{0n} = 1$

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-12

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 : avec interaction

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

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

Présentation du transparent facultative

Les variables TESTEUR et ORIGINE sont **qualitatives** et ce sont donc des **facteurs**, que l'on indicera par i (pour le facteur testeur : T1 et T2) et par j (pour le facteur origine : Afrique, Hongrie, Maroc). La répétition r représente l'individu r de testeur i et d'origine j , n_{ij} le nombre d'individus de testeur i et d'origine j, et N le nombre total d'individus.

Construction du modèle : analyse de variance à deux facteurs avec interaction

— μ : moyenne générale

— α_i : effet du niveau i du facteur A : testeur

— β_j : effet du niveau j du facteur B : origine

— γ_{ij} : effet du niveau ij de l'interaction entre les facteurs A et B

— ϵ_{ijr} : erreur résiduelle (ce qui n'est pas expliqué par le modèle).

Le paramètre θ_0 représente la moyenne μ.
Le paramètre θ_p représente l'effet du testeur p pour p variant de 1 à 2 et l'effet de l'origine pour p variant de 3 à 5 et l'effet de chaque couple d'interaction pour p variant de 6 à 11 ($I \times J = 6$ couples (i,j) d'interaction). Les x_{pijr} prennent les valeurs 1 ou 0 selon que les niveaux i, j et ij correspondent ou non au paramètre θ_p.

Par exemple, θ_6 correspond à l'interaction entre le testeur T1 (facteur testeur i=1) et l'Afrique (facteur origine j=1).

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-13

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

$$\begin{aligned} \mathrm{Y_n} &= \alpha + \beta.\mathrm{Z_n} + \varepsilon_\mathrm{n} \quad \mathrm{n} = 1 \cdots \mathrm{N} \\ \mathrm{Y_n} &= 1.\theta_0 + \mathrm{Z_n}\theta_1 + \varepsilon_\mathrm{n} \quad \mathrm{n} = 1 \cdots \mathrm{N} \end{aligned}$$

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

$$\begin{aligned} \mathrm{Y_n} &= \beta_0 + \beta 1.\mathrm{Z1_n} + \beta 2.\mathrm{Z2_n} + \beta 3.\mathrm{Z3_n} + \epsilon_\mathrm{n} \\ \mathrm{Y_n} &= 1.\theta_0 + \mathrm{Z1_n}\theta_1 + \mathrm{Z2_n}\theta_2 + \mathrm{Z3_n}\theta_3 + \epsilon_\mathrm{n} \end{aligned}$$

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

$$\mathrm{Y_n} = 1.\theta_0 + \mathrm{Z1_n}\theta_1 + \mathrm{Z2_n}\theta_2 + \mathrm{Z3_n}\theta_3 + \mathrm{Z4_n}\theta_4 + \mathrm{Z5_n}\theta_5 + \epsilon_\mathrm{n}$$

Remarque : Sous la forme de TP, on s'appuiera sur la connaissance pratique de la régression et de l'analyse de la variance pour donner des écritures possibles de modèles. On mettra en évidence la linéarité en les paramètres à l'aide de l'écriture avec les paramètres θ_p.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-14

Exemple 3 : Écriture du modèle relatif à l'exemple tension artérielle

Toutes les variables sont **quantitatives,** on se trouve donc dans le cadre d'une régression à un **régresseur**. La tension est la *variable expliquée* et l'âge est la *variable explicative*.
Le modèle choisi est une droite où les inconnues sont α et β. L'ordonnée à l'origine est le coefficient α et la pente de la droite est le coefficient β. On écrira le modèle équivalent avec $\theta_0 = \alpha$ et $\theta_1 = \beta$. On insistera sur la linéarité en les paramètres θ_p.

$$Y_n = \sum_{p=0}^{C-1} x_{pn}\,\theta_p + \epsilon_n$$

Dans le cadre d'une régression simple, on a :

— n (individu n, observation n)
— paramètre p (0 à C-1 = 1) à estimer
— $x_{1n} = Z_n$
— $x_{0n} = 1$

Exemple 4 : Écriture du modèle relatif à l'exemple rendement du blé

Toutes les variables sont **quantitatives,** on se trouve donc dans le cadre d'une régression à 3 **régresseurs**. Le rendement est la *variable expliquée*. Les inconnues à estimer sont les coefficients α et $\beta_1,\ \beta_2,\ \beta_3$ ou encore $\theta_0\ et\ \theta_1, \theta_2, \theta_3$.

$$Y_n = \sum_{p=0}^{C-1} x_{pn}\,\theta_p + \epsilon_n$$

Dans le cadre d'une régression à 3 régresseurs, on a :

— n (individu n, observation n)
— paramètre p (0 à C-1 = 3) à estimer
— $x_{pn} = Zp_n$ p = 1 à 3
— $x_{0n} = 1$

Exemple 5 : Écriture du modèle relatif à l'exemple moutarde

On présentera directement l'écriture sous forme de θ_p.

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)

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

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

Remarque : On expliquera la construction du modèle progressivement.
Si on considère que la variable dose est **qualitative**, on construit alors un modèle d'analyse de la variance à un **facteur** et on note i le niveau du facteur traitement (dose), i variant de 1 à 5.

formulation courante :

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

Construction du modèle : analyse de variance à un facteur

— μ : moyenne générale

— α_i : effet du niveau i du facteur A : traitement (dose)

— ϵ_{ir} : erreur résiduelle (ce qui n'est pas expliqué par le modèle).

Tableau des correspondances entre le niveau i du facteur A et les paramètres θ_p :

Niveau i du facteur A	θ_1	θ_2	θ_3	θ_4	θ_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

Le paramètre θ_0 représente la moyenne μ.
Mais, comparer les niveaux du facteur traitement (dose) entre eux, c'est ne pas tenir compte de la croissance de la dose. L'idée consiste à décomposer le facteur A en une régression linéaire simple ou quadratique ou polynomiale en fonction de la quantité de la variable **quantitative** dose.

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

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

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

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

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

En Régression linéaire simple : on utilise la quantité du régresseur dose Z_i pour mieux expliquer l'effet du niveau i du facteur traitement (dose)

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

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

Tableau de liaison entre les différentes écritures de modèle :

paramètres	paramètres
μ	θ_0
β	θ_1
α'_1	θ_2
$\vdots$	$\vdots$
α'_5	θ_6

Construction du modèle : analyse de covariance

— μ : moyenne générale

— Z_i : régresseur quantité de la dose

— α'_i : écart entre α_i et la régression $(\beta\,.\,Z_i)$

En Régression linéaire quadratique : on utilise la quantité du régresseur dose Z_i et du régresseur Z_i^2 pour mieux expliquer l'effet du niveau i du facteur traitement (dose)

$$Y_{ir} = \mu + (\beta_2 \,.\, Z_i^2 + \beta_1 \,.\, Z_i + \alpha_i'') + \epsilon_{ir}$$

$$\begin{cases} \mathrm{Y_{1r}} & = 1.\theta_0 + \mathrm{Z_1^2}.\theta_1 + \mathrm{Z_1}.\theta_2 + 1.\theta_3 + 0.\theta_4 + 0.\theta_5 + 0.\theta_6 + 0.\theta_7 + \epsilon_{1\mathrm{r}} \\ \vdots & \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \mathrm{Y_{5r}} & = 1.\theta_0 + \mathrm{Z_5^2}.\theta_1 + \mathrm{Z_5}.\theta_2 + 0.\theta_3 + 0.\theta_4 + 0.\theta_5 + 0.\theta_6 + 1.\theta_7 + \epsilon_{5\mathrm{r}} \end{cases}$$

Tableau de liaison entre les différentes écritures de modèle :

paramètres	paramètres
μ	θ_0
β_1	θ_1
β_2	θ_2
α_1''	θ_3
$\vdots$	$\vdots$
α_5''	θ_7

Construction du modèle : analyse de covariance

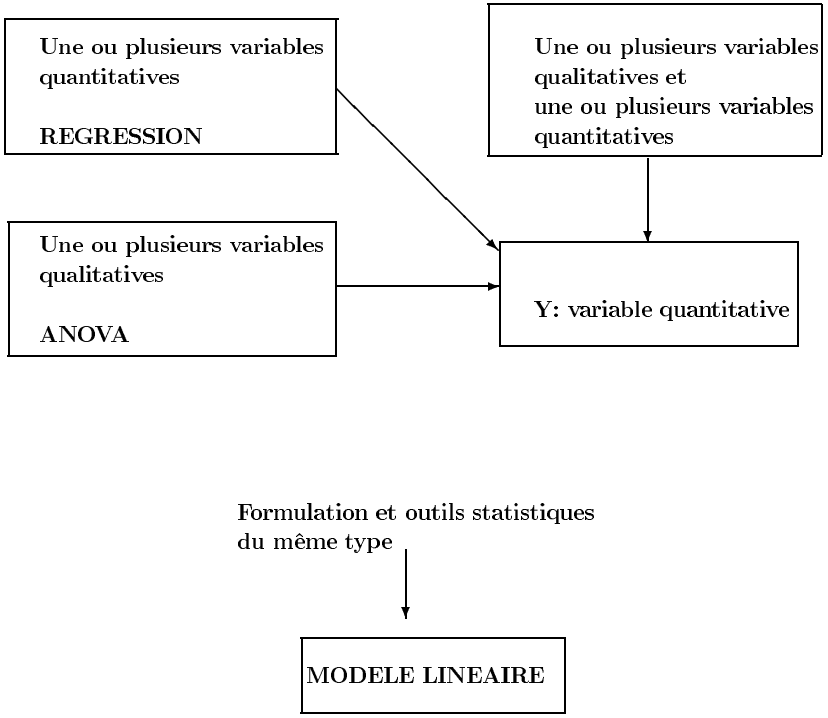
— μ : moyenne générale

— Z_i : régresseur quantité de la dose

— α_i'': écart entre α_i et la régression $(\beta_2 \,.\, Z_i^2 + \beta_1 \,.\, Z_i)$

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-16

Synthèse sur les Exemples



Schéma

Replacer les différentes méthodes dans un schéma unique et montrer la similarité d'approche.

Utilisation d'un même modèle.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-17

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

Tous les exemples vus précédemment peuvent s'écrire sous la forme d'un même modèle statistique :

$$Y_n = x_{0n} \,.\, \theta_0 + x_{1n} \,.\, \theta_1 + x_{2n} \,.\, \theta_2 + \ldots + x_{pn} \,.\, \theta_p + \ldots + \epsilon_n$$

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

ou encore

$$Y_n = \mu_n + \epsilon_n$$

Comme μ_n est **linéaire en ses paramètres** (inconnus), ce modèle est appelé le **modèle linéaire.**

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

Adopter le modèle linéaire c'est se donner un cadre général : à savoir

— poser comme principe que la composante systématique du modèle est linéaire dans ses paramètres :

— $\mu_n = \sum x_{pn}.\theta_p$

— faire des **postulats** sur la partie résiduelle ϵ_n du modèle

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-19

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 :

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

Un modèle sera dit **linéaire**, s'il est linéaire en ses paramètres. Ce type de modèle ne se réduit pas à la droite. En fait, il s'agit de toutes les fonctions pouvant s'écrire sous la forme :

$$E(Y_n) = \mu_n = \sum f(Zv) \, . \, \theta_p$$

où les $f(Zv)$ sont des fonctions connues des prédicteurs Zv : c'est à dire ne dépendant pas des paramètres θ_p inconnus. C'est encore une combinaison linéaire des paramètres θ_p.

Exercice : Faire rechercher aux stagiaires les x_{pn} correspondant aux exemples du transparent

— **Exemple 1** : on a $x_{0n} = 1$, $x_{1n} = Z1_n$ et $x_{2n} = (Z1_n)^2$

— **Exemple 2** : on a $x_{0n} = 1$, $x_{1n} = \log(Z1_n)$ et $x_{2n} = e^{Z2_n}$

— **Exemple 3** : on a $x_{0n} = 1$, $x_{1n} =?$ et $x_{2n} = \frac{1}{Z1_n + \theta_1}$.

Les exemples 1 et 2 sont linéaires en $\theta_1, \theta_2, \theta_3$ même s'ils comportent des fonctions *puissance, exponentielle* ou *logarithme*. Ce sont les prédicteurs qui sont les arguments de ces fonctions.

Dans l'exemple 3, x_{2n} est fonction du paramètre θ_1, il est de plus impossible d'expliciter x_{1n}. Le modèle n'est pas linéaire

Exercices

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. $\cdots$

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-21

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 : $E(\epsilon_{\mathrm{n}}) = 0$;

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

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

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

Choisir le modèle linéaire, c'est faire un certain nombre de **postulats** concernant la loi conjointe de ϵ_n.

On parle de **postulats** plutôt que d'*hypothèses* puisqu'on les suppose vrais au départ et qu'ils sont invérifiables.

Certaines méthodes descriptives et statistiques permettent parfois de soupçonner la non-réalisation de ces postulats. Elles seront détaillées au chapitre 7.

Remarque : une seule notation permet de résumer les 4 postulats sur le vecteur de ϵ_n

$$\epsilon_n \; iid \sim \mathcal{N}(0, \sigma^2)$$

(iid : indépendant et identiquement distribué)

Remarque : Les postulats 1 et 2 permettent d'écrire

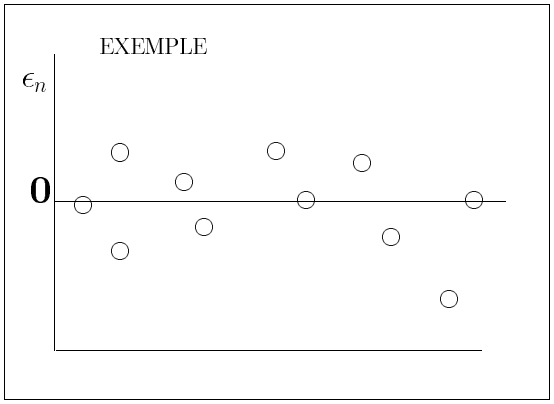
$$E(Y_n) = \mu_n = \sum x_{pn} \, . \, \theta_p$$

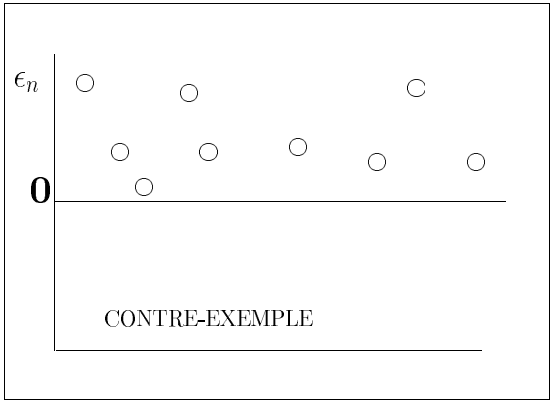
et que

$$\mathrm{Var}\,(Y_n) = \sigma^2$$

Si le postulat 4 est aussi vérifié alors Y_n suit aussi une loi Normale $\mathcal{N}(E(Y_n), \sigma^2)$.

Postulats du modèle linéaire
$E(\epsilon_n) = 0$





Dire que l'espérance des ϵ_n est nulle, c'est dire que le modèle posé est correct, que toutes les sources de variabilité (autre que expérimentale) ont été prises en compte.

C'est aussi dire que la partie linéaire du modèle ($\mu_n = \sum x_{pn} \, . \, \theta_p$) rend compte du phénomène étudié. En effet, si $E(\epsilon_n) = 0$, l'espérance des Y_n est égale à μ_n:

$$Y_n = \mu_n + \epsilon_n$$

soit $E(Y_n) = E(\mu_n + \epsilon_n) = \mu_n + E(\epsilon_n) = \mu_n + 0 = \mu_n$

La figure présente deux situations hypothétiques et invérifiables ; en effet, on ne disposera après l'estimation des paramètres du modèle que des résidus estimés $\widehat{\varepsilon}_n$ (estimations des erreurs ε_n) qui par construction seront tels que $E(\widehat{\varepsilon}_n) = 0$.

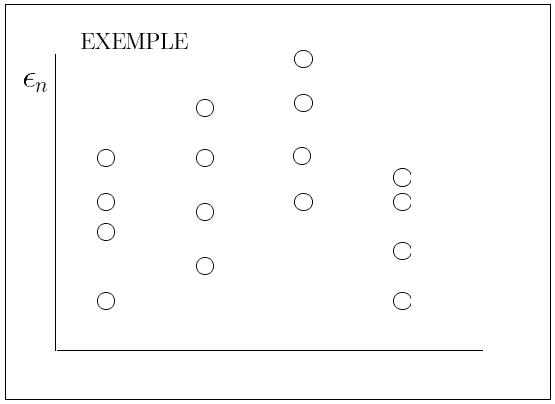
Dans le premier exemple, les ϵ_n sont centrées sur 0 avec autant de valeurs au dessus et en dessous de 0.

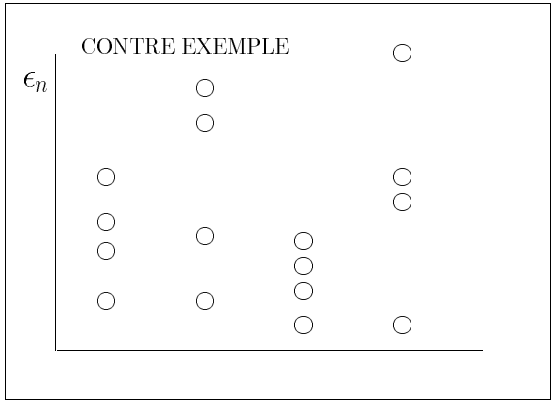
Dans le deuxième exemple, l'ensemble des valeurs est au dessus de 0. Il reste une ou des sources de variabilité.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-23

Postulats du modèle linéaire

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

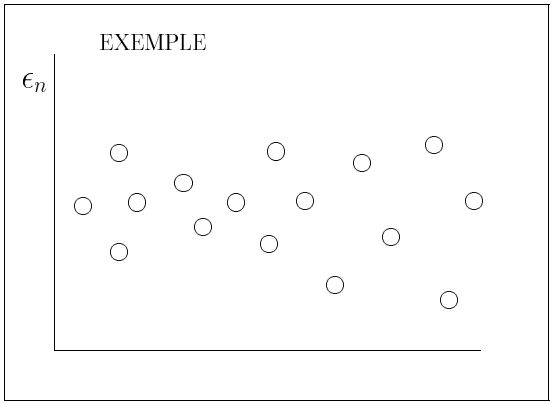


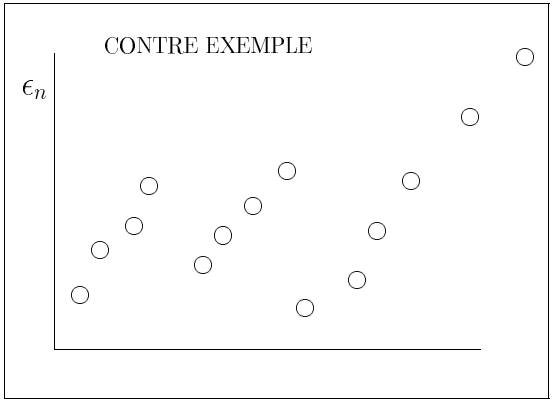


La variance des termes d'erreur est constante $\forall n$. C'est le postulat d'homoscédasticité.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-24

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





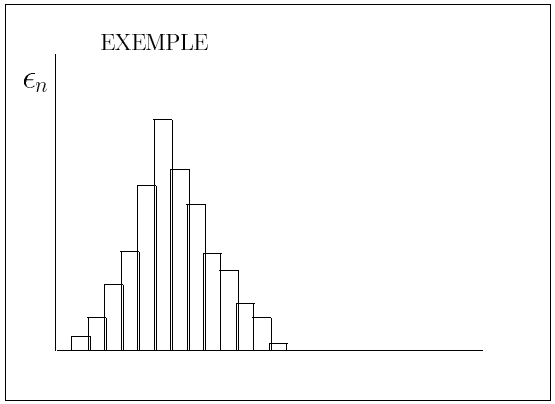
Postulat d'indépendance.

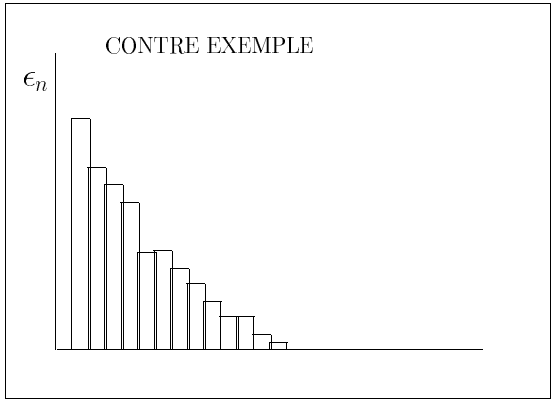
Ce postulat est vérifié quand chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. Ce postulat peut ne pas être vérifié dans le cas de séries temporelles ou spatiales.

Compte tenu des postulats 2 et 3, la matrice de variance-covariance des résidus s'écrit :

$$\begin{pmatrix} \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

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





Les ϵ_n suivent des lois gaussiennes ou normales. Il permet de faire des tests :

C'est le postulat le moins important et on peut s'en passer lorsque l'on dispose d'un jeu de données important.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-26

Liens avec les autres modules de FPSTAT

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.

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

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

Dans certaines circonstances et pour certains jeux de données, on peut sortir du modèle linéaire.

— si Y_n n'est pas quantitatif, mais plutôt du type (0,1), on utilisera le Modèle Linéaire Généralisé (GLM).

— 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 appartient à la famille des lois exponentielles, on entre dans le cadre du Modèle Linéaire Généralisé (GLM).

— si les ϵ_n ont des variances différentes, on peut utiliser comme technique d'estimation les *moindres carrés pondérés*, ou transformer les variables pour stabiliser la variance ;

— si la matrice de variance-covariance est non diagonale, on entre dans le cadre du Modèle Mixte.

FPSTAT INRA 1997 I-Introduction au Modèle Linéaire-27