Les ε_i vérifient les 4 postulats :

P1. $\mathrm{E}(\varepsilon_i) = 0$

P2. $\mathrm{var}(\varepsilon_i) = \sigma^2$ ne dépend pas de i

P3. les ε_i sont indépendantes

P4. les ε_i suivent des lois gaussiennes

(Les postulats du modèle linéaire sont présentés juste après la définition du Modèle Linéaire statistique :)

- Les conditions d'utilisation du modèle linéaire général sont précisées par les 4 postulats suivants, tous relatifs au vecteur des erreurs ε_i.

- Nous parlons ici de ***postulats*** plutôt que d'***hypothèses*** puisque ces 4 hypothèses sont invérifiables. Il faudrait en effet disposer des vrais termes d'erreur alors que ne sont accessibles que les estimations de ces termes d'erreurs.

  Certaines méthodes descriptives permettent parfois de soupçonner la non-réalisation de ces postulats sur les données elles-mêmes. C'est ce que nous allons présenter tout de suite pour chacun de ces 4 postulats. Soulignons que cette vérification n'est pas suffisante et qu'une vérification plus approfondie doit être réalisée sur les résidus . L'étude de ces résidus sera détaillée ultérieurement au cours des journées consacrées à la régression et à l'analyse de variance.

  Remarque formateur : une seule notation permet de résumer les 4 postulats sur le vecteur des erreurs

$$\varepsilon_i \quad iid \sim N\left(0, \sigma^2\right)$$

$$\downarrow$$

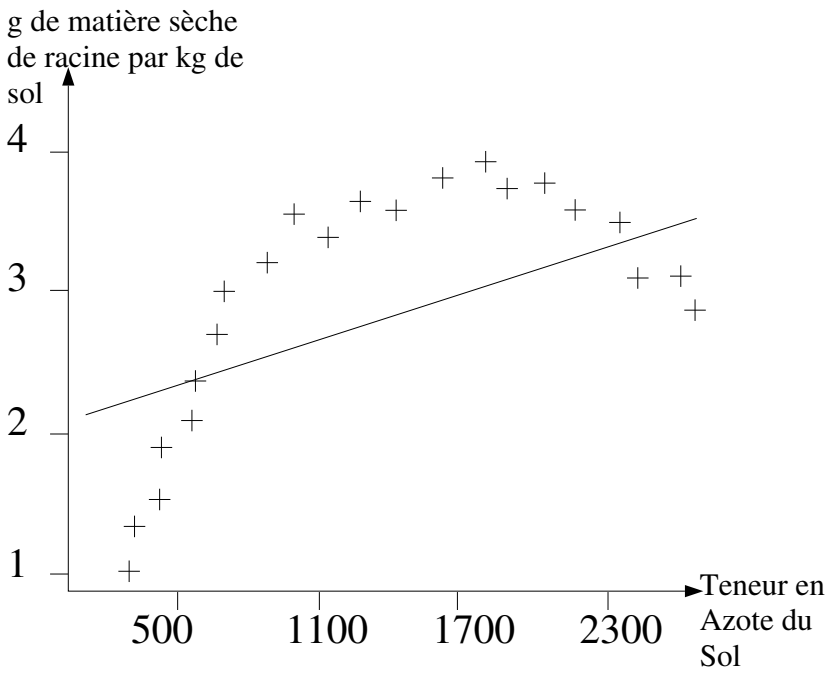
indépendantes et identiquement distribuées

POSTULAT 1:

$$\mathrm{E}(\varepsilon_i) = 0$$

Les erreurs sont centrées, leur espérance est nulle

Contre-exemple en régression linéaire



D'après Maury et Rivoire, 1963

- En clair, cela veut dire que le modèle posé est correct, que l'on n'a pas oublié un terme pertinent.

- Lors d'une expérience sur une graminée, la fléole, l'estimation de la matière sèche de racine produite en fonction de la teneur en azote du sol conduit au graphique suivant.

  Dans un tel cas, les prévisions données par le modèle seraient sous estimées pour les valeurs moyennes de la teneur en azote du sol et sur estimées pour les valeurs extrêmes.

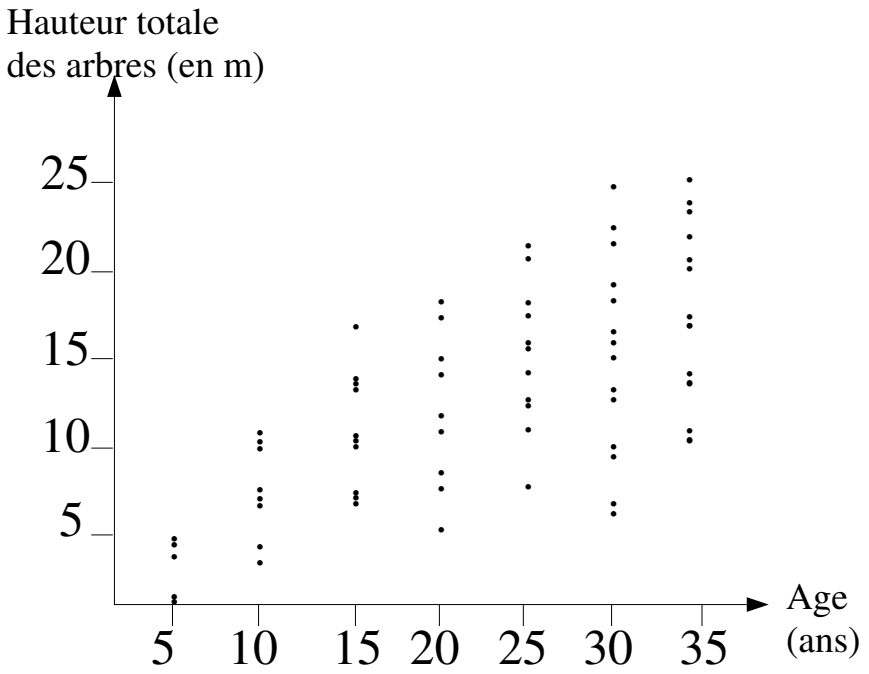
  Le modèle linéaire choisi est incorrect.

- En **analyse de variance**, une circonstance gênante est celle où un facteur de variation parasite a été oublié et où le dispositif a été fait de telle sorte que certains des ses effets se confondent avec ceux des traitements. Citons, par exemple, le cas d'une comparaison de variétés de tomate en serre où certaines variétés seraient systématiquement placées près du système d'irrigation et d'autres plus loin. La perturbation causée par un éventuel effet d'irrigation non contrôlé pourrait conduire à des conclusions fausses. Le modèle d'analyse de variance à un facteur serait incorrect.

POSTULAT 2:

$$\text{var}\,(\varepsilon_i) = \sigma^2 \quad \text{ne dépend pas de } i$$

Contre-exemple en régression linéaire



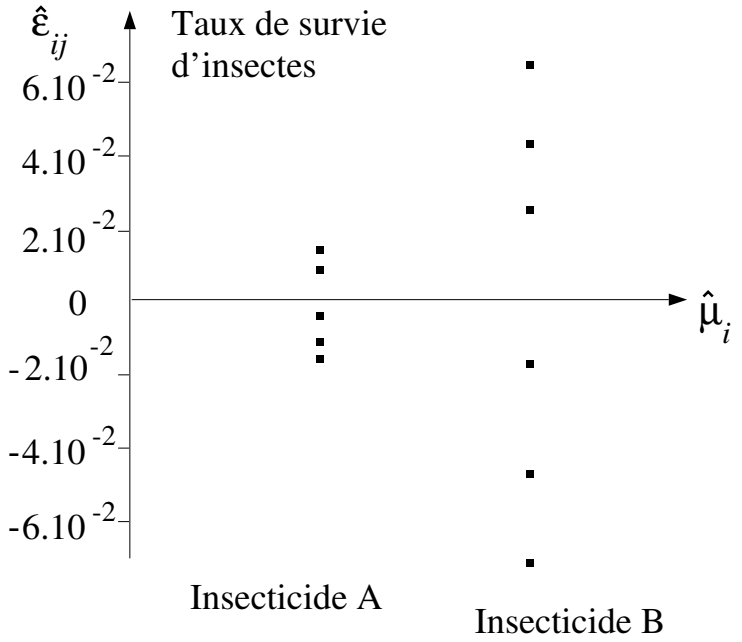
- La variance des termes d'erreur est constante. C'est le postulat d'**homoscédasticité**.

- On étudie la croissance en hauteur de l'Epicea commun dans le Jura entre 5 et 35 ans. On observe que la variance des hauteurs observées croît avec l'âge (i) et donc avec la hauteur moyenne observée.

POSTULAT 2:

$$\text{var}(\varepsilon_i) = \sigma^2 \quad \text{ne dépend pas de } i$$

Contre-exemple en analyse de variance



- La variance des termes d'erreur est constante. C'est le postulat d'**homoscédasticité**.

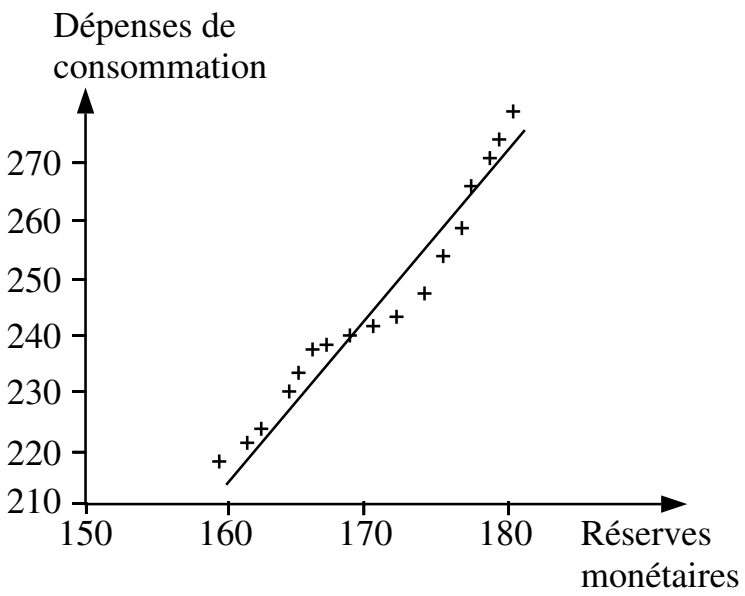
- On étudie le taux de survie d'insectes à 2 insecticides A et B. On fait différentes répétitions de l'expérience et on obtient le graphique suivant des résidus $\widehat{\varepsilon_{ij}}$ en fonction des moyennes $\widehat{\mu_i}$ des 2 insecticides.

- Le produit A est plus efficace que le produit B. Le taux de survie au produit A est proche de zéro et peu variable d'une répétition à l'autre. Le taux de survie au produit B est plus élevé mais la variabilité des réponses observées est aussi plus grande.

POSTULAT 3:

les composantes ε_i sont indépendantes

Contre-exemple en régression linéaire



D'après Chatterjee et Price, 1977

- Ce postulat est vérifié quand chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. Par contre, dans les problèmes où le temps joue un rôle important, il est plus difficilement vérifié.

- On étudie la régression des dépenses de consommation aux Etats-Unis en fonction des réserves monétaires du pays (ces réserves monétaires ont été calculées entre le premier trimestre 1952 et le dernier trimestre 1956).

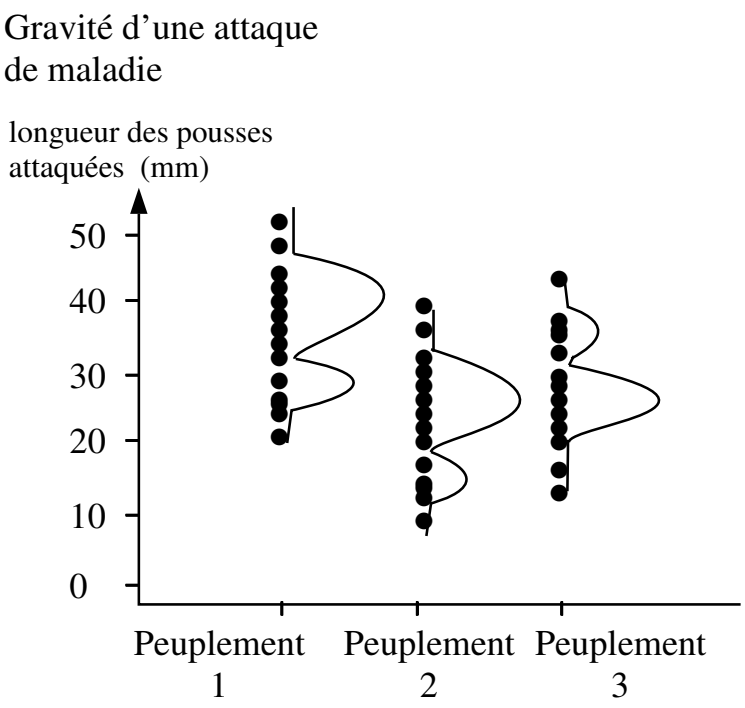
- Dans ce contre-exemple, la variable explicative est elle-même une fonction du temps et il y a une certaine rémanence ou inertie du phénomène étudié.

- En **analyse de variance**, le postulat est plus difficilement vérifié lorsque les différentes modalités d'un facteur sont étudiées sur les mêmes individus à des temps t différents (ex : étude de croissance en fonction de l'âge).

POSTULAT 4:

les ε_i suivent des lois gaussiennes ou normales

Contre-exemple en analyse de variance



- C'est le postulat le moins important, en particulier on peut s'en passer lorsqu'on dispose d'un lot de données important.

  Remarque : ce postulat ne concerne donc pas uniquement le nuage global de points mais aussi chacun des niveaux du facteur concerné.

- La gravité d'attaque d'une maladie cryptogamique appelée Rouille est étudiée sur différents peuplements de pin sylvestre. La variable observée est la longueur des pousses touchées par une attaque.

  La distribution des notes de gravité apparaît bimodale et disymétrique dans chaque peuplement.