Partie 5. Etude des relations entre variables

Etude de la liaison entre variables: corrélation vs régression

Rappels :

Corrélation : Etablir l'existence d'un lien entre 2 VA (X, Y) jouant des rôles **symétriques**

→ Pas d'hypothèse sur sens relation (~~VD~~, ~~VI~~) ou de cause à effet.

→ On s'intéresse seulement à la dispersion du nuage de point, pas à sa pente.

Ex : Est-ce que la quantité d'azote (X) et le rendement en maïs (Y) varient ensemble?
Si oui, avec quelle intensité et dans quel sens ?

Régression : Trouver le meilleur modèle permettant de décrire et **quantifier l'influence** d'1 ou plusieurs **VI** (X_i, fixes et/ou aléatoires) **sur** les variations d'1 **VD** (Y, aléatoire).

→ lien **asymétrique** : il y a un sens dans la relation, et c'est ce qu'on cherche à modéliser.

→ On s'intéresse à la dispersion ET à la **pente** du nuage de point.

*Ex : Est-ce que le rendement en maïs (Y) varie linéairement **en fonction de** la quantité d'azote (X)?*
Si oui, de combien?

Modèles de régression

Qu'est ce qu'un modèle ?

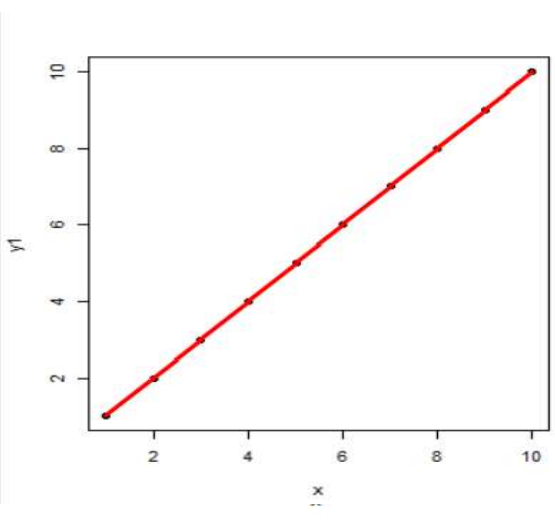
« Simplification » de la réalité sous forme d'une équation mathématique traduisant la relation existant entre les variables d'intérêt *(ex: y = f(x)).*

Quel modèle (équation) correspond le mieux à mes données ?

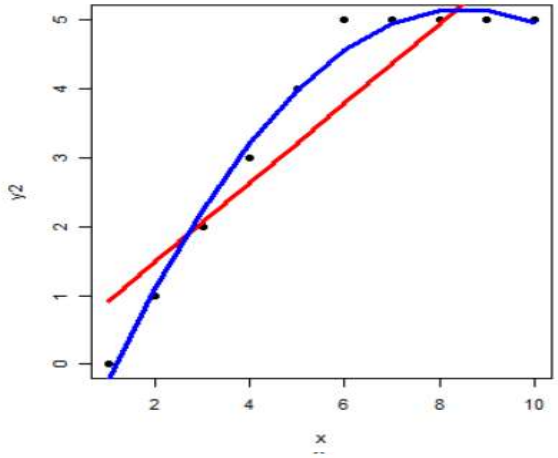
Importance de la visualisation des données.

R^2 = coefficient de détermination; varie entre [0, 1] et mesure de la qualité de l'ajustement du modèle aux observations.

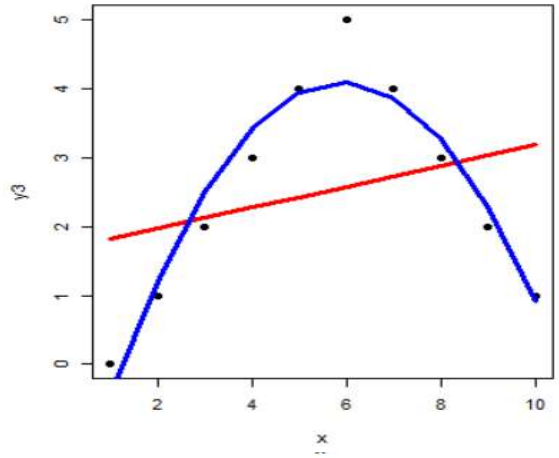
Modèle linéaire, R^2 = 1



Modèle linéaire, R^2 = 0.84
Modèle non linéaire, R^2 = 0.98



Modèle linéaire, R^2 = 0.08
Modèle non linéaire, R^2 = 0.93



Modèles de régression

Pourquoi utiliser la régression ?

- **Comprendre / décrire** la relation entre une VD (Y) et 1 ou plusieurs VI (X_j)

→ Ajuster le meilleur modèle et en estimer les paramètres à partir des données.

Si les variations de Y sont étudiées en fonction de :

- 1 seule variable X → régression simple
- Plusieurs variables X_j → régression multiple

- **Prédire**, à partir du modèle, la valeur attendue (= estimée) de Y pour 1 ou plusieurs nouvelles valeurs de X_j (= non observées dans l'échantillon).

Ex:

*- **Comprendre** comment le rendement en maïs (Y) varie en fonction de la quantité d'azote apportée (X_1), la température (X_2), les précipitations (X_3) et la densité de ravageurs (X_4)?*

*- **Prédire** le rendement en maïs à partir de valeurs particulières de quantité d'azote, température, précipitation et densité de ravageurs (nouvelles valeurs de X_1 à X_4).*

Partie 5. Etude des relations entre variables

La régression linéaire simple

Les données : n individus statistiques →1 VD quantitative continue (Y) et 1 VI quantitative (X).

Objectif : Evaluer l'existence d'1 lien <u>linéaire</u> entre Y et X.

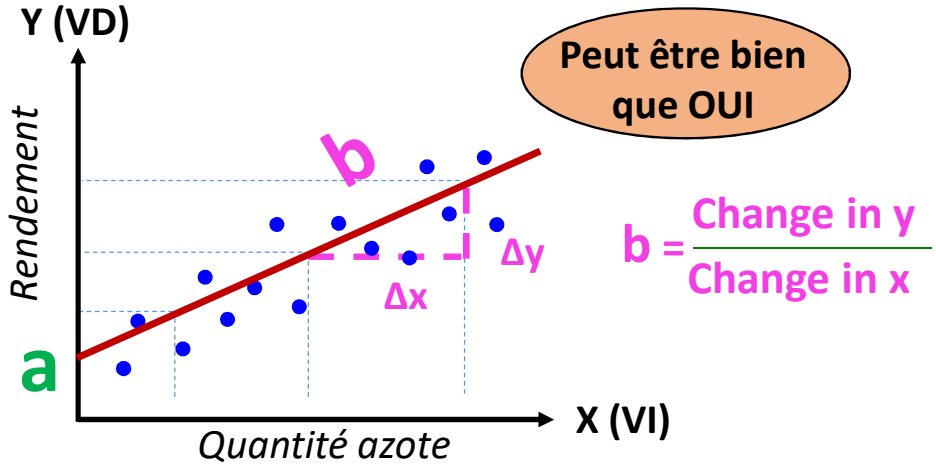
Relation résumée par 1 droite d'équation :

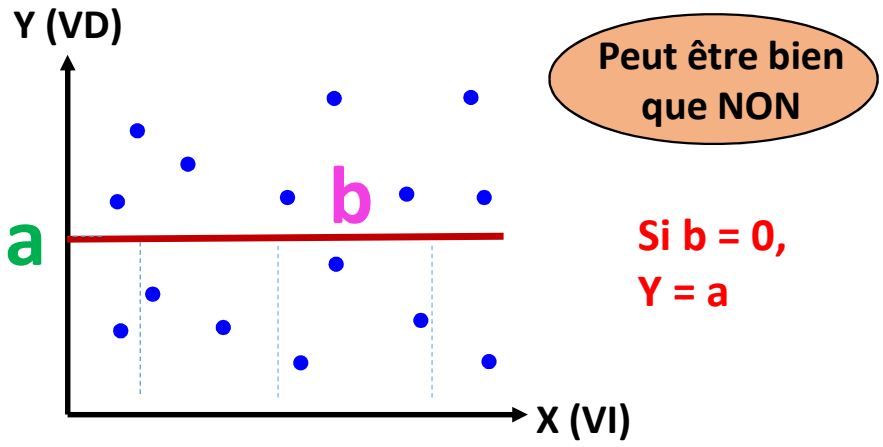
$$y = a + b\,x$$

Ordonnée à l'origine = Intercept
(valeur de Y quand X vaut 0)

Pente (= coefficient directeur)

Ex : Est ce que le rendement (y) peut être expliqué/prédit par la quantité d'azote apportée (x) ?





La régression linéaire simple

Principe :

Déterminer l'équation de la droite de régression qui passe au plus près de l'ensemble des observations (points x_i, y_i), en la faisant pivoter sur le point moyen ($\overline{x}$, $\overline{y}$).

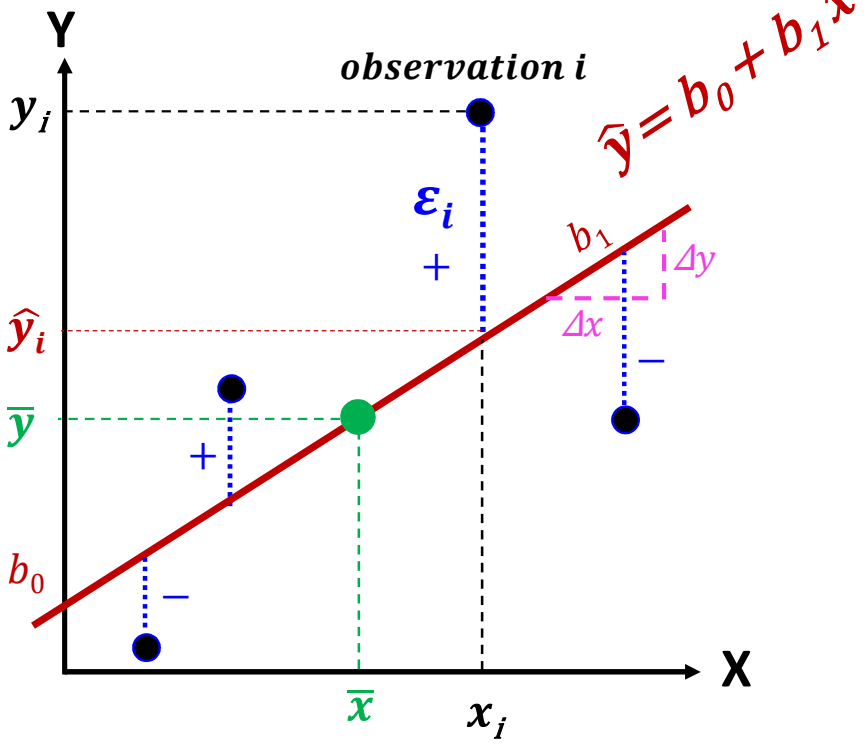
Ecriture usuelle droite : $\hat{y} = b_0 + b_1 x$

Résidu (erreur) = écart entre chaque observation i (x_i, y_i) et la valeur de y prédite ($\hat{y}_i$) par le modèle (cad la droite) pour un x_i donné.

$$\varepsilon_i = y_i - \hat{y}_i$$

La meilleure droite sera celle qui minimise la somme des carrés des résidus ε_i (= méthode des moindres carrés).

Ecriture modèle : $y_i = \hat{y}_i + \varepsilon_i = b_0 + b_1 x_i + \varepsilon_i$



La régression linéaire simple

Evaluation de la qualité globale du modèle :

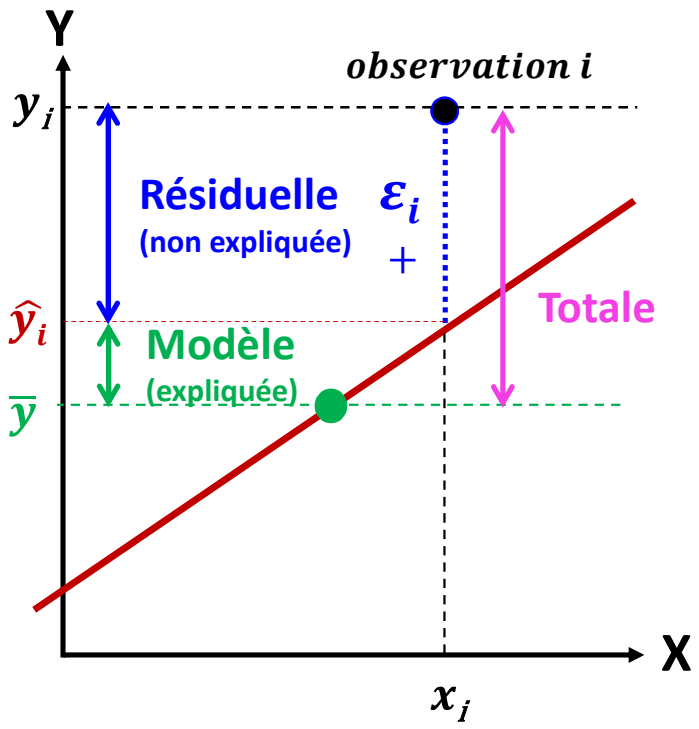
La **variabilité totale de Y** (écart entre y_i et $\overline{y}$) peut se décomposer en 2 parties:

- **Expliquée par le modèle** (écart entre $\overline{y}$ et $\widehat{y_i}$) : pour chaque valeur de x_i le modèle prédit $\widehat{y_i}$
- **Non expliquée** (aléatoire) = **Résiduelle** (écart entre $\widehat{y_i}$ et y_i) : ≠ prédiction *vs* observation

Quelle part (%) de la variabilité de Y s'explique par X ?

$$SCE_T = SCE_M + SCE_R$$

Variabilité	SCE	ddl	Carré moyen	Test	Proba. crit.
Totale	SCE_T	$n-1$			
Due au modèle	SCE_M	1	$CM_F=SCE_F$	$F_{obs}=CM_F/CM_R$	p_c
Résiduelle	SCE_R	$n-2$	$CM_R=SCE_R/(n-2)$		



✓ **Test de Fisher :** Plus F_{obs} est grand (variance Modèle >> Résiduelle), meilleur est l'ajustement du modèle aux données.

✓ **Coefficient de détermination multiple** $R^2 = SCE_M / SCE_T$

= Proportion de la variance totale de Y expliquée par le modèle.

La régression linéaire *(simple ou multiple)*

Estimation des coefficients de la régression (paramètres inconnus, non observables) **:**

Régression linéaire <u>simple</u> : $y_i = b_0 + b_1 x_{i1} + \varepsilon_i$

Régression linéaire <u>multiple</u> : $y_i = b_0 + \underbrace{b_1 x_{i1} + b_j x_{ij} + \cdots + b_p x_{ip}}_{\text{combinaison linéaire (somme) de plusieurs VI } (X_j)} + \varepsilon_i$

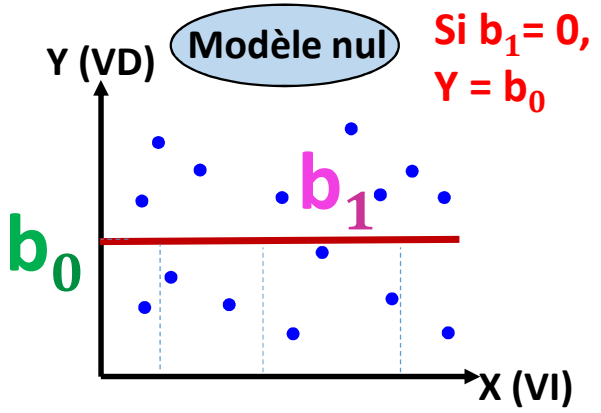
➔ **Dans tous les cas on a :**

- 1 **constante** b_0 = **ordonnée à l'origine**
  → Pas vraiment d'intérêt pour évaluer s'il existe 1 relation linéaire entre variables.
- 1 **coefficient** de **pente** b_j par VI (j = 1 à p) = variation de Y pour une ↗ d'1 unité de x_j (lorsque toutes les autres VI sont maintenues constantes).
  → Renseigne sur la vitesse d'évolution de la variable y en fonction de la variable x_j, mais ne présume pas de la <u>significativité</u> de la relation (doit la tester : Student).

+ 1 **résidu** ε_i = variation de Y non expliqué par les VI du modèle (erreur)

La régression linéaire *(simple ou multiple)*

2 niveaux de test de la significativité du modèle :



1. Le test global : *Le modèle est-il pertinent/informatif ?*

Statistique de test : F de Fisher.

Hypothèses :

- **H_0 :** $b_1 = b_i = \ldots = b_p = 0$ → « toutes les pentes = 0, aucune des VI (X_j) n'influence Y»
  (= modèle nul avec seulement b_0)
- **H_1 :** $b_1 \neq b_i \neq \ldots \neq b_p \neq 0$ → « au moins 1 des pentes ≠ 0, au moins 1 VI joue »

Interprétation :

- Si on conclue **H_0** → on s'arrête là !

(Soit les VI sont mal choisies et le modèle n'est pas bon, soit la taille de l'échantillon est insuffisante)

- Si on conclue **H_1** → on peut interpréter le R^2 pour apprécier la qualité du modèle.
  (régression simple = « Multiple R-squared», régression multiple = «Adjusted R-squared»)

La régression linéaire *(simple ou multiple)*

2 niveaux de test de la significativité du modèle :

2. Le test de significativité de chaque coefficient : *La VI X_i est-elle utile ?*

Statistique de test : t de Student.

Hypothèses :

H_0 : $b_i = 0$ → test de « la pente $b_i = 0$» pour chaque VI

H_1 : $b_i \neq 0$ → cette VI est significativement liée linéairement à Y

Interprétation :

- Si on conclue **H_0** → cette VI n'est pas liée linéairement à Y (variations de X sans effet sur Y)

(en régression multiple, on retire les VI non significatives du modèle pour optimiser sa capacité de prédiction)

- Si on conclue **H_1** → valeur du coefficient renseigne sur la taille et sens de l'effet de X sur Y

Remarque : *Quand on compare des modèles, on retiendra celui qui maximise le R^2 tout en limitant le nombre de VI → principe de parcimonie (cherche a minimiser le critère AIC ou le BIC)*

La régression linéaire simple

Conditions d'application :

Pour pouvoir utiliser la régression linéaire **4 conditions** doivent être satisfaites :

1. La relation entre X et Y est **linéaire** (au moins grossièrement).
   → Vérification graphique : pas de pattern autre que linéaire (ex: relation en U non linéaire)

2. Les observations sont **indépendantes** (résidus non corrélés) : Très important.
   → C'est le plan d'échantillonnage qui renseigne sur cette condition.

   Si les données proviennent d'individus statistiques différents et qu'il n'y a pas d'échantillonnage dans l'espace ou le temps, elles sont généralement indépendantes.

La régression linéaire simple

3. Les résidus doivent suivre une loi normale centrée sur 0 (i.e. sur droite de régression)

4. Les résidus doivent avoir une variance homogène (homoscédasticité)

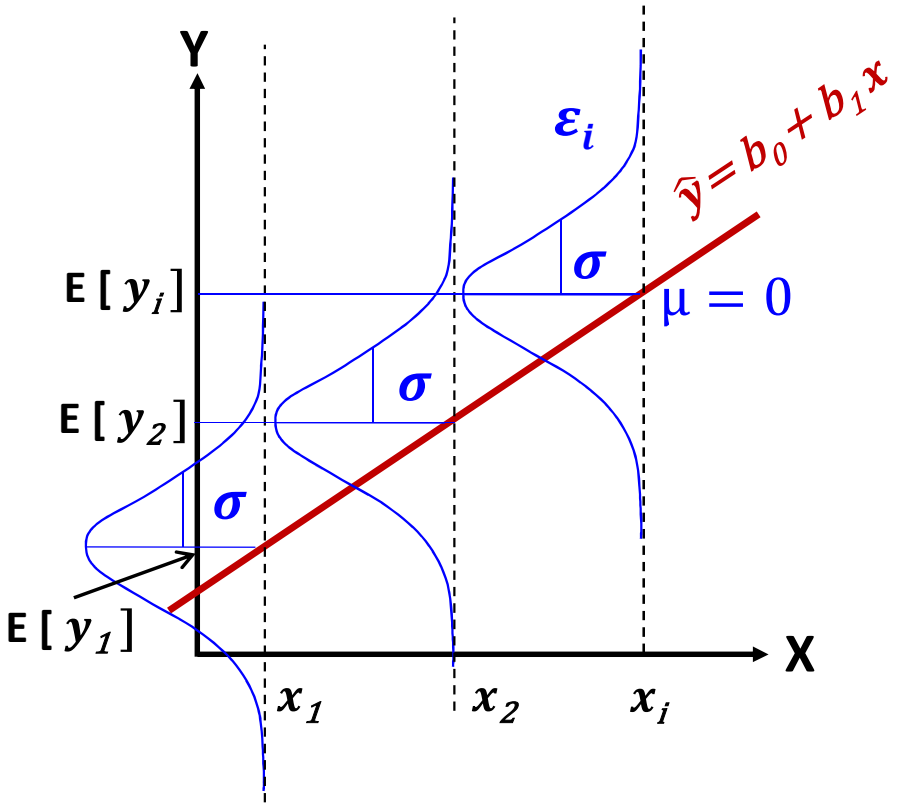
→ Vérification graphique et tests.

Explication :

Si on répète un grand nombre de fois les mesures de Y pour 1 x_i donné, on s'attend en moyenne à ce que les y_i se trouve sur la droite de régression → espérance ***E [y_i].***

Autour de cette valeur ***E [y_i],*** *on aura une distribution Normale des résidus* $\boldsymbol{\varepsilon_i}$ (**μ** centrée sur 0 –sur la droite- et $\boldsymbol{\sigma}$) dans laquelle se trouvera la valeur réellement observé de y_i .

Cette distribution des résidus N(**0**,$\boldsymbol{\sigma}$) doit être la homogène pour toutes les valeurs de x_i .



La régression linéaire *(simple ou multiple)*

Flux de travail :

1. Visualiser les données.
2. Créer un modèle.
3. Tester les conditions d'application de base.
4. Interpréter les résultats du modèle.