Corrélation - régression simple

G. Marot-Briend
guillemette.marot@univ-lille.fr

2021-2022

Plan

Introduction

Langage courant :
Corrélation = liaison entre deux variables quelque soit leur nature.

Sens statistique

- Corrélation : évaluation de la liaison entre deux variables quantitatives (le plus souvent, liaisons essentiellement linéaires)
- Régression : méthode permettant de proposer un modèle mathématique pour expliquer les relations entre les observations.

Introduction

Langage courant :
Corrélation = liaison entre deux variables quelque soit leur nature.

Sens statistique

- Corrélation : évaluation de la liaison entre deux variables quantitatives (le plus souvent, liaisons essentiellement linéaires)
- Régression : méthode permettant de proposer un modèle mathématique pour expliquer les relations entre les observations.

Problèmes ne relevant pas de la corrélation :

- liaison entre deux variable qualitatives $\Rightarrow \chi^2$
- liaison entre une variable qualitative et une variable quantitative $\Rightarrow$ comparaison de plusieurs moyennes, ANOVA

Notations

On considère n individus sur lesquels on mesure X et Y deux variables quantitatives discrètes ou continues.

Pour chaque individu i $(1 \leq i \leq n)$, on dispose d'un couple d'observations (x_i, y_i) qui représente les valeurs prises par X et Y pour l'individu i.

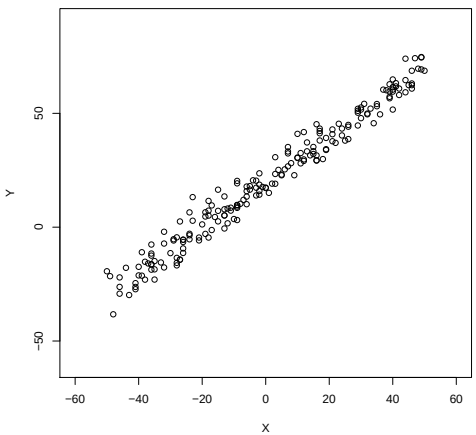
On cherche à "expliquer" Y en fonction de X, c'est-à-dire à exprimer une dépendance fonctionnelle de Y comme fonction de X du type $Y = f(X)$.
On appelle Y la variable à expliquer, X la variable explicative.

Représentations graphiques

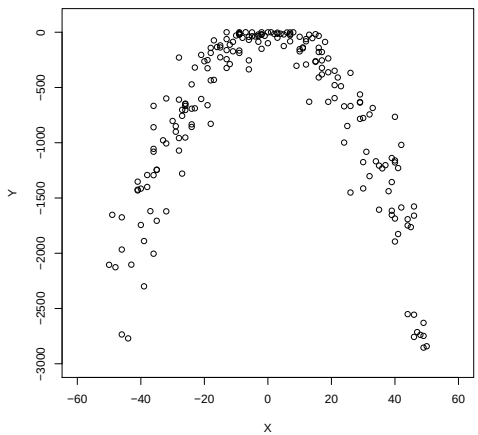
Graphique pour représenter deux variables quantitatives

$\Rightarrow$ nuage de points

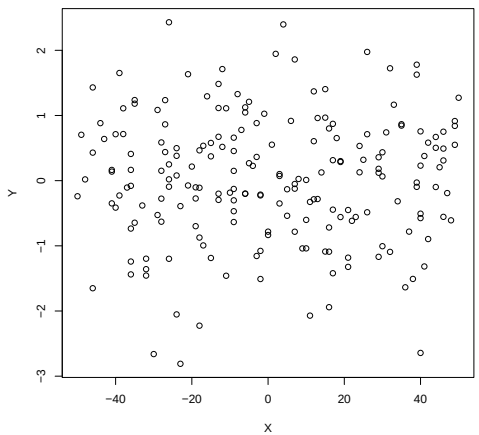
1ère étape de toute analyse de liaison : apprécier la forme de la relation entre les deux variables



liaison linéaire



liaison polynomiale



pas de liaison

Plan

Covariance

Covariance :

Mesure de la variation simultanée de deux variables aléatoires. La covariance permet d'évaluer l'importance et le sens de cette variation.

$$\sigma_{XY} = \text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

- si les variables sont liées, la covariance est importante.
- une covariance peut être positive, négative ou nulle.

Covariance

Covariance empirique :

$$\text{cov}_{ech}(x, y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

Remarques :

- $\text{cov}(X, Y) = \text{cov}(Y, X)$
- $\text{cov}(aX, Y) = a\text{cov}(X, Y) = a\text{cov}(Y, X)$
- $\text{cov}(X, X) = \text{Var}(X)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X, Y)$

Coefficient de corrélation

Coefficient de corrélation linéaire noté ρ_{XY}

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Interprétation du coefficient de corrélation :

ρ mesure la relation entre deux variables quantitatives X et Y, ρ est toujours compris entre -1 et 1.

- si $\rho = 0$, les variations des variables X et Y sont indépendantes.
- si $\rho > 0$, les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
- si $\rho < 0$, les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.

La liaison est d'autant plus forte que $|\rho|$ est proche de 1.

Coefficient de corrélation

Indépendance et corrélation :

- si X et Y sont indépendantes, alors $E(X, Y) = E(X)E(Y)$ $\Rightarrow \rho_{XY} = 0$
- si $\rho_{XY} = 0$ et X et Y sont distribuées normalement, alors X et Y sont indépendantes.

Coefficient de corrélation

Le coefficient de corrélation mesure de façon symétrique la relation entre les deux variables, sans notion de contrôle sur l'une des deux variables : $\rho_{XY} = \rho_{YX}$

Estimation du coefficient de corrélation

Bravais Pearson :

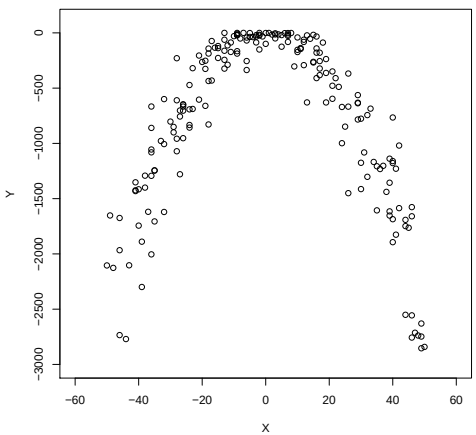
$$r(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

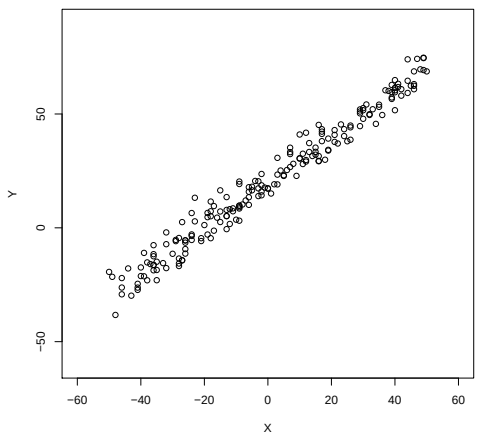
Coefficient de corrélation

Remarques :

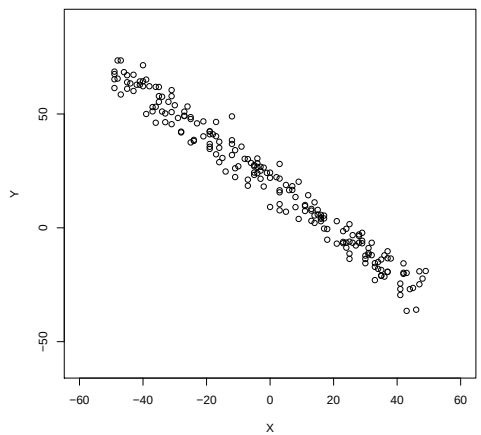
- r est très sensible aux valeurs extrêmes.
- On peut avoir une liaison même si $r(x, y) = 0$;
  r mesure seulement le caractère linéaire d'une liaison.



$r(x, y) = 0$



$r(x, y) > 0$



$r(x, y) < 0$

Coefficient de corrélation

Exemple : Fréquence cardiaque maximale (FCM)

On souhaite étudier une relation éventuelle entre l'âge d'un individu, notée X et sa FCM, variable notée Y

Individu k	Age x_k	FCM y_k
1	40	187
2	36	195
3	51	180
4	49	190
5	47	185
6	51	183
7	32	195
8	55	185
9	55	189
10	23	201
11	49	189
12	52	185
13	35	195

Questions :

1. Calculer $\bar{x}$ et $\bar{y}$
2. Calculer $\text{cov}_{ech}(x, y)$, $s^2_{ech}(x)$ et $s^2_{ech}(y)$
3. Calculer $r(x, y)$

Correction

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{575}{13}$$
$$= 44,23$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{2459}{13}$$
$$= 189,15$$

$$\text{cov}_{ech}(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$
$$= \frac{108157}{13} - \frac{575}{13}\frac{2459}{13}$$
$$= -46,65$$

Correction

$$
\begin{aligned}
s^2_{ech}(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \\
&= \frac{26641}{13} - \left(\frac{575}{13}\right)^2 \\
&= 92,95
\end{aligned}
$$

$$
\begin{aligned}
s^2_{ech}(y) &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 \\
&= \frac{465551}{13} - \left(\frac{2459}{13}\right)^2 \\
&= 32,44
\end{aligned}
$$

Correction

$$r(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

$$\begin{aligned} r(x,y) &= \frac{\mathrm{cov}_{ech}(x,y)}{s_{ech}(x)s_{ech}(y)} \\ &= \frac{-46,65}{\sqrt{92,95.32,34}} \\ &= -0,85 \end{aligned}$$

Coefficient de corrélation

Test du coefficient de corrélation

Principe et hypothèses :

Si $\rho = 0$ alors il n'y a pas de liaison linéaire entre X et Y
Si $\rho \neq 0$ alors il y a une liaison linéaire entre X et Y

$$\begin{cases} \mathbf{H_0} : \rho = 0 \\ \mathbf{H_1} : \rho \neq 0 \end{cases}$$

Conditions de validité du test :

Test valide pour une distribution binormale du couple de variables aléatoires (X,Y). La binormalité correspond à une distribution normale, en chaque point de X, de la variable Y et vice versa.

En pratique, on se contente souvent de vérifier la normalité des distributions de X ou Y et le caractère monotone de leur relation.

Test du coefficient de corrélation

Statistique de test : sous H_O,

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim St_{n-2}$$

Région critique :

$$W =]-\infty; -t_{n-2\ ddl,1-\alpha/2}] \cup [t_{n-2\ ddl,1-\alpha/2}; +\infty[$$

Décision :
Si $t \in W$ alors on rejette H_0 au risque de première espèce α.
Il existe une liaison linéaire entre X et Y.

Test du coefficient de corrélation

Exemple : FCM et âge

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$
$$= \frac{-0,85\sqrt{13-2}}{\sqrt{1-0,85^2}}$$
$$t = -5,35$$

$$t_{11} = 2,201$$

Sous réserve de validité du test (binormalité), il existe une liaison linéaire entre l'âge et la fréquence cardiaque maximale.

Test du coefficient de corrélation

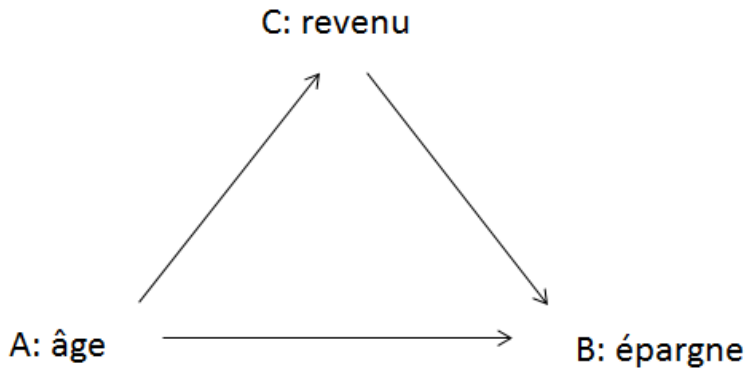
Remarques :

- La loi de R est aussi tabulée et permet de calculer des seuils de significativité pour n donné.
  Ex : au risque 5%, pour $n = 30$, on déclare qu'une liaison est significative si $|r| > 0.36$.

- Le test est robuste mais si les conditions ne sont clairement pas vérifiées, alors on utilisera un test non paramétrique.

Corrélations partielles

En pratique, il arrive fréquemment que la liaison observée entre 2 paramètres soit en fait due aux variations d'un 3$^{\text{ème}}$ paramètre, appelé facteur de confusion.



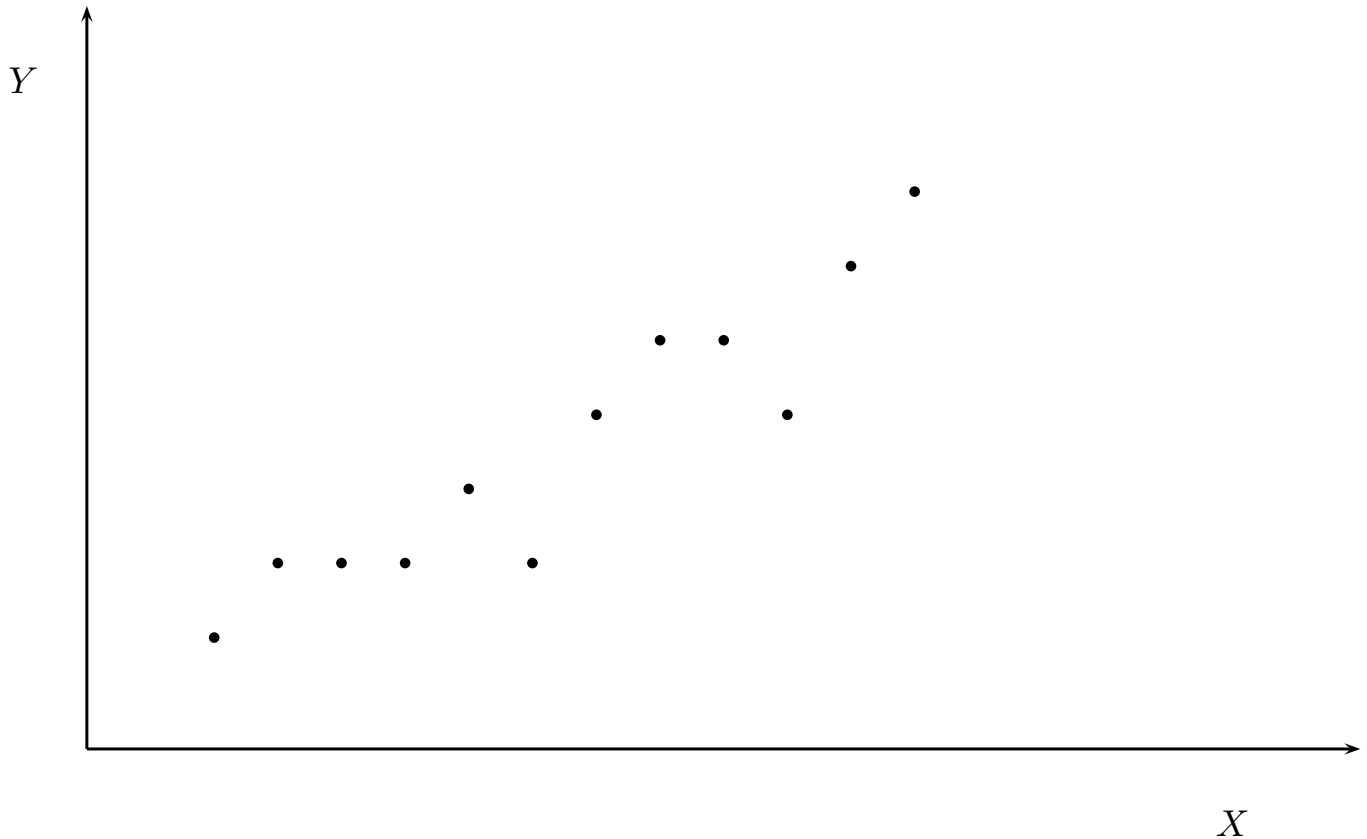
Exemple : mesure de l'épargne annuelle en fonction de l'âge des salariés d'une entreprise.

Corrélation partielle

$$\rho_{AB/C} = \frac{\rho_{AB} - \rho_{AC}\rho_{BC}}{\sqrt{(1-\rho_{AC}^2)(1-\rho_{BC}^2)}}$$

Plan

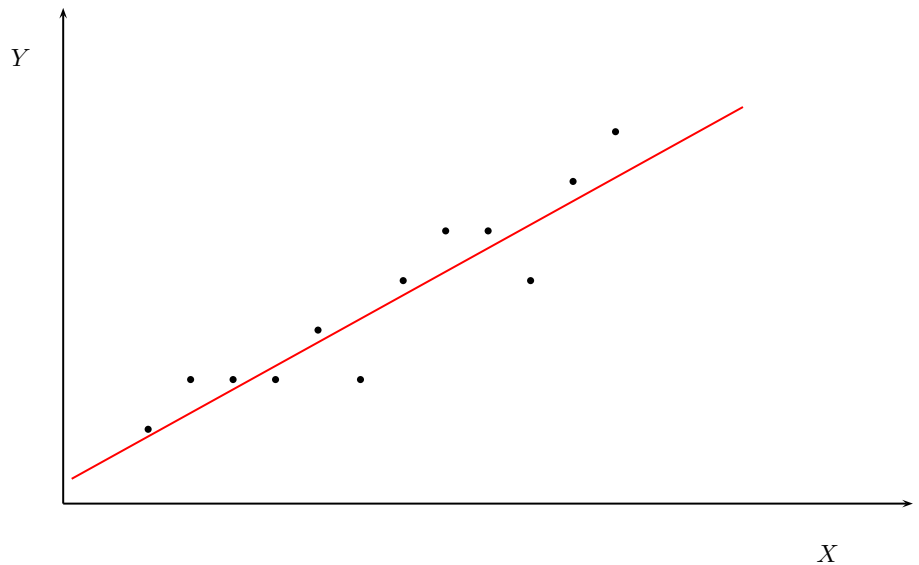
Régression linéaire simple



Régression linéaire simple

La régression linéaire simple consiste à proposer une droite pour expliquer une v.a. quantitative par une autre

$$Y = f(X) + \epsilon$$



Régression linéaire simple

Modèle de régression

$$Y = \alpha X + \beta + \epsilon$$

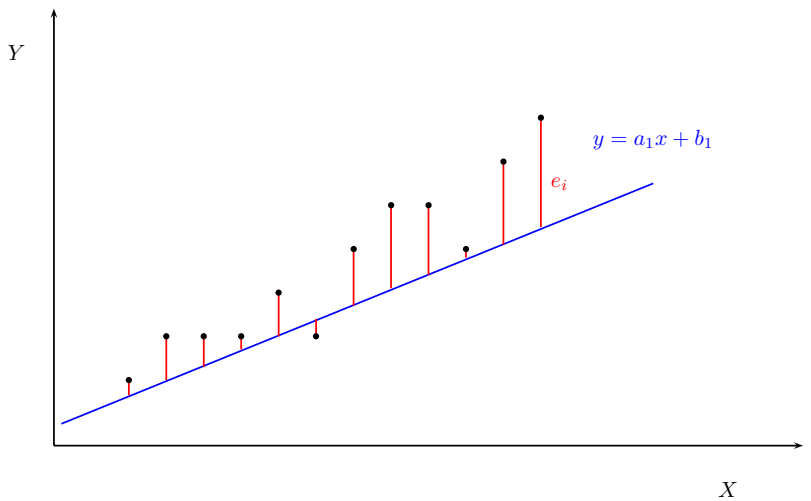
Hypothèses : $\forall i, j$

- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (normalité des erreurs)
- $E[\epsilon_i] = 0$ (erreurs centrées)
- $V[\epsilon_i] = \sigma^2$ (homoscédasticité des erreurs)
- $E[\epsilon_i \epsilon_j]_{i \neq j} = 0$ (erreurs indépendantes - non corrélées)

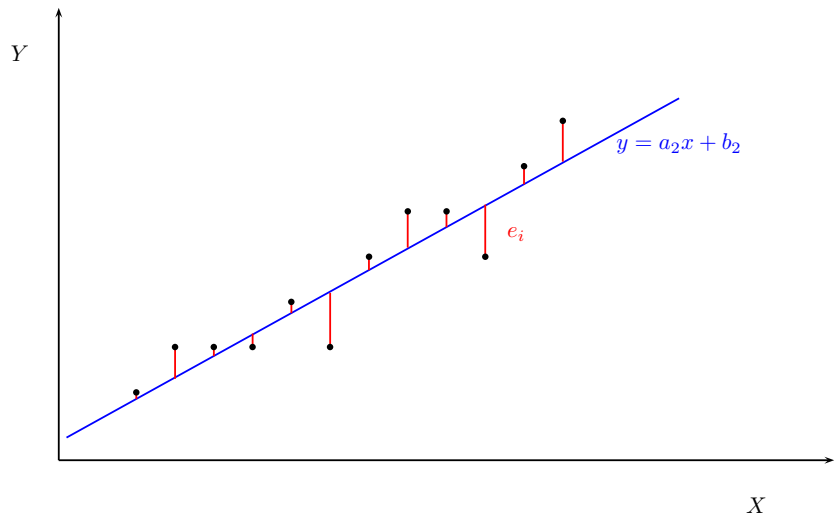
Régression linéaire simple

A partir des données observées dans un échantillon,

$$y_i = ax_i + b + e_i$$



erreur importante



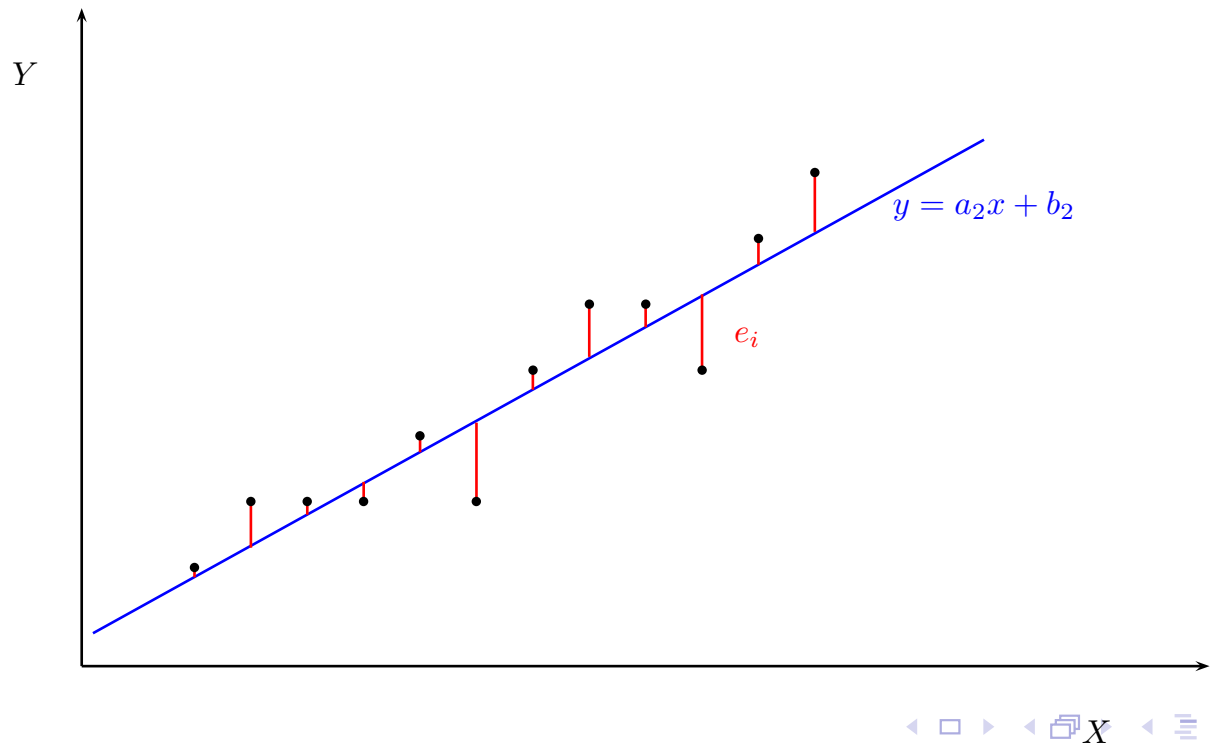
erreur minimisée

Régression linéaire simple

Méthode des Moindres Carrés Ordinaires (MCO)

Minimiser la somme des carrés des écarts

$$\varphi(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$



Régression linéaire simple

Solutions de la minimisation :

$$\hat{a} = \frac{\text{cov}(x, y)}{\text{Var}(x)}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Remarque :

la pente de la droite de régression peut être déduite du coefficient de corrélation r

$$\hat{a} = r\frac{s_y}{s_x}$$

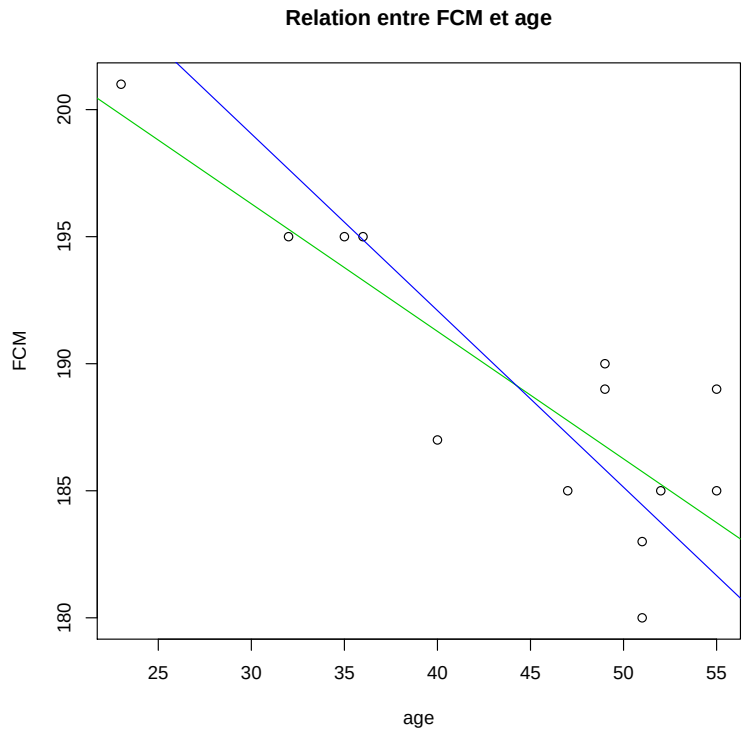
Régression linéaire simple

Exercice :

Individu k	Age x_k	FCM y_k
1	40	187
2	36	195
3	51	180
4	49	190
5	47	185
6	51	183
7	32	195
8	55	185
9	55	189
10	23	201
11	49	189
12	52	185
13	35	195

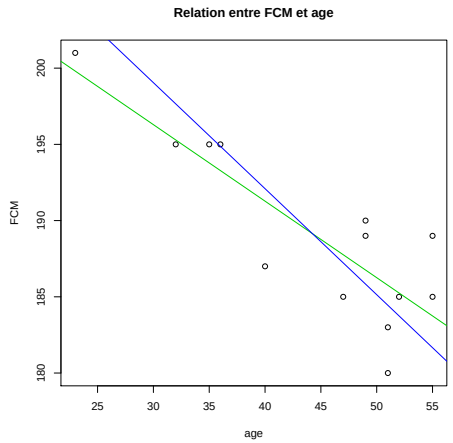
1. expliquer la FCM par l'âge
2. Tracer la droite des MCO sur le nuage de points
3. expliquer l'âge par la FCM
4. Tracer cette deuxième droite des MCO sur le nuage de points

Régression linéaire simple



Les deux droites des MCO sont en général distinctes, elles se coupent toujours au centre de gravité du nuage $(\bar{x}, \bar{y})$.

Régression linéaire simple



L'angle entre ces deux droites donne une mesure de la dépendance entre les variables X et Y : plus cet angle est ouvert, moins la liaison est forte :

- les deux droites de MCO sont confondues $\Longleftrightarrow$ il y a liaison linéaire exacte entre X et Y
- les deux droites de MCO sont perpendiculaires si les deux variables X et Y sont non corrélées.

Régression linéaire simple

Prévision avec la droite des MCO

Si x^* est une nouvelle valeur de X, on prédira la valeur $\hat{y^*}$ de Y donnée par la relation

$$\hat{y^*} = \hat{a}x^* + \hat{b}$$

- s'assurer de la qualité de l'ajustement avant de donner des prévisions
- une prévision d'une valeur de Y n'a de sens que pour des valeurs de X proches de celles utilisées pour déterminer $\hat{a}$ et $\hat{b}$

Prévision avec la droite des MCO

Démarche

1. calcul de la droite des MCO
2. validation du modèle $\Rightarrow$ étude des résidus et détection des valeurs aberrantes et influentes
3. qualité de l'ajustement $\Rightarrow$ décomposition de la variance, coefficient de détermination et test significativité globale
4. qualité de prédiction (PRESS) et prédiction

Prévision avec la droite des MCO

Etude des résidus

On appelle valeur ajustée de la $i^{\text{ème}}$ observation de la variable Y l'approximation

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

On appelle résidu e_i, l'erreur observée que l'on commet en approchant y_i par $\hat{y}_i$: $e_i = y_i - \hat{y}_i$

Ne pas confondre erreurs non observables et résidus.

Etude des résidus

Validité du modèle

- vérifier la normalité des résidus
- vérifier que les résidus ne contiennent pas d'information structurée
- vérifier que les résidus ne sont pas auto-corrélés entre eux

cf. cours spécifique MISO

Qualité de l'ajustement

Equation d'analyse de la variance

$$y_i - \bar{y} = (\widehat{y_i} - \bar{y}) + (y_i - \widehat{y_i})$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Somme des carrés totale SCT}} = \underbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2}_{\text{Somme des carrés expliquée SCE}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}_{\text{Somme des carrés résiduelle SCR}}$$

Qualité de l'ajustement

Equation d'analyse de la variance

$$y_i - \bar{y} = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i)$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Somme des carrés totale SCT}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\text{Somme des carrés expliquée SCE}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\text{Somme des carrés résiduelle SCR}}$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\text{Variance expliquée}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\text{Variance résiduelle}}$$

Qualité de l'ajustement

Coefficient de détermination

Part de la variance de y expliquée par la relation $\hat{y} = \hat{a}x + \hat{b}$

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)}$$

Dans le cas d'un ajustement linéaire, on peut montrer que $R^2 = r^2(x, y)$ (où r est le coefficient de corrélation linéaire)

- $R^2 \in [0, 1]$
- Plus R est proche de 1, plus le modèle explique correctement la variabilité de Y.

Plan

Conclusion

Croisement de deux variables quantitatives

Représentation graphique (nuage de points)

Coefficient de corrélation

- Calcul de l'indicateur statistique
- Test de nullité du coefficient de corrélation

Régression linéaire

- Estimation des coefficients
- Validité du modèle (Etude des résidus et des observations influentes)
- Qualité d'ajustement (R^2, significativité globale)
- Prédiction