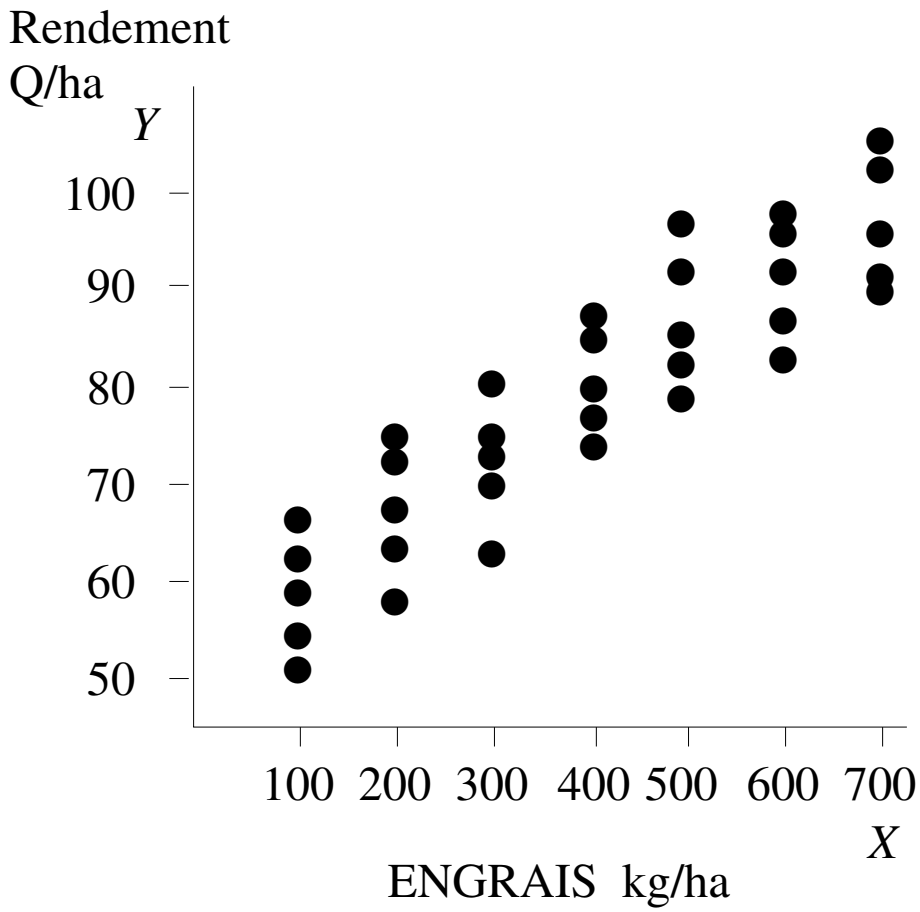
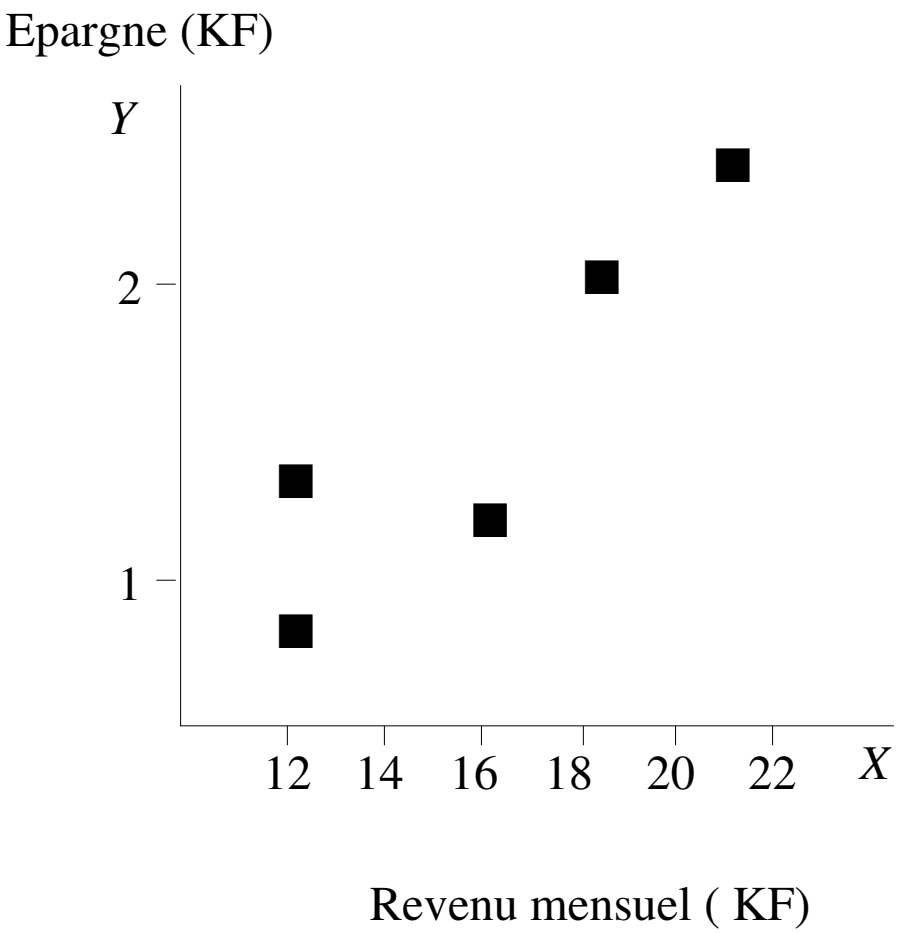
BLE


Rendement
Q/ha
Y
100
90
80
70
60
50
100 200 300 400 500 600 700
X
ENGRAIS kg/ha

- Présenter le domaine d'application de la régression et son intérêt. Dire que la régression concerne la relation entre 2 variables.
- Par exemple, on cherche à étudier l'influence de la dose d'engrais sur le rendement de blé. Dans cette expérience, on a noté le rendement de 35 parcelles traitées par diverses doses d'engrais (7 doses, 5 parcelles ou 5 répétitions par dose)
- Souligner la différence entre les 2 types de variables : dose d'engrais et rendement.
  - Les 7 doses utilisées sont fixées préalablement par l'expérimentateur. La variable dose est dite non-aléatoire (ou régresseur, ou variable indépendante). On la symbolise par X ; valeurs sur l'axe des abscisses.
  - Le rendement varie d'une parcelle à l'autre pour la même dose (dose = 100 kg/ha $\rightarrow$ le rendement varie entre 48 et 62 Q=Quintaux/ha). C'est une variable aléatoire, symbolisée par Y; les valeurs sont portées sur l'axe des ordonnées.
  - Montrer que le rendement augmente à peu près linéairement avec la quantité d'engrais. Comme Y dépend de X, on appellera Y variable dépendante.
- La régression doit permettre :
  - **de décrire l'importance de l'effet "dose" en ajustant au mieux une courbe à travers l'ensemble des points ;**
  - **d'exprimer le rendement en fonction de la dose (l'inverse n'a pas de sens pour l'expérimentateur) ;**
  - **éventuellement prédire Y pour X donné.**

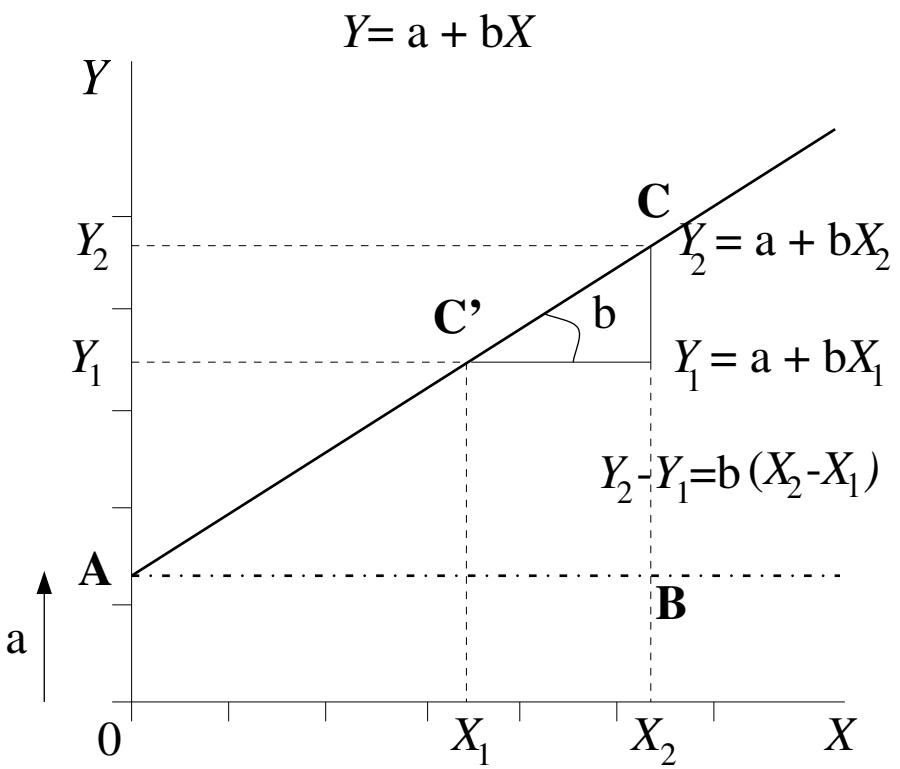
FAMILLES



- Dans l'exemple précédent, l'augmentation de Y est causée par l'augmentation de X. On tire des conclusions de cause à effet.
- Une telle conclusion n'est pas toujours possible.

  Exemple : un échantillon de 5 familles représentatives d'une catégorie donnée ; on note pour chacune d'elles le revenu et le montant mensuel de l'épargne.
- Les valeurs de l'épargne ou du revenu n'ont pas été fixées à priori ; les 2 variables sont toutes deux aléatoires.
- Le choix entre celle qui joue le rôle de X et celle qui joue le rôle de Y est flou.

  On peut rechercher la relation qui lie l'épargne en fonction du revenu (ou l'inverse), mais on ne peut pas conclure à l'existence d'un lien de causalité.
- Lorsque X et Y sont aléatoires, l'accroissement de Y, bien que parallèle à l'accroissement de X, peut dépendre d'une autre variable non étudiée, dont l'influence s'exerce aussi bien sur X que sur Y.
- La relation la plus simple entre 2 variables est celle où Y est relié linéairement à X ; dire qu'une relation est linéaire sous-entend que la variation de l'une est proportionnelle à l'autre.



$$b=(Y_2-Y_1)/(X_2-X_1)$$

(Transparents 3 et 4 facultatifs selon le niveau des stagiaires ; à présenter ou à faire au tableau)

- La relation la plus simple est la relation linéaire ; on rappellera les notions fondamentales de l'équation d'une droite.
- Une relation linéaire entre 2 variables Y et X fait correspondre à la variable X la variable Y telle que $Y = a + bX$ où a et b sont des constantes.
- Le graphe de cette fonction est une droite.
- Quand $X = 0, Y = a$; a est l'ordonnée à l'origine, c'est à dire l'ordonnée du point d'intersection de la droite et de l'axe des ordonnées.
- Signification de la constante b.

  Soient 2 points C et C' de coordonnées $C(X2, Y2)$ et $C'(X1, Y1)$

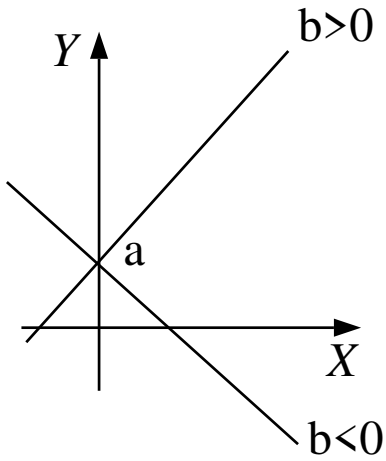
  $$Y2 = a + bX2 \quad et \quad Y1 = a + bX1$$

  par différence : $Y2 - Y1$ on obtient :

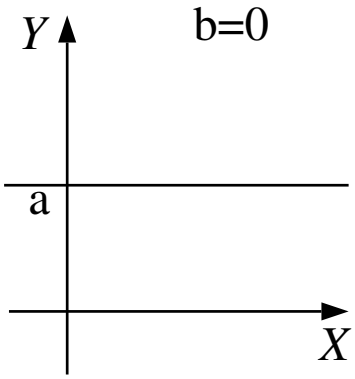
  $$b = (Y2 - Y1)/(X2 - X1)$$

  b = différence des ordonnées / différence des abscisses
- b est appelé coefficient directeur de la droite. Il représente la pente ou la tangente de l'angle en C' (côté opposé/côté adjacent) = tangente de$B\widehat{A}C$. Il traduit donc la valeur de l'angle formé par la droite et l'axe des abscisses, donc l'inclinaison de la droite sur cet axe.

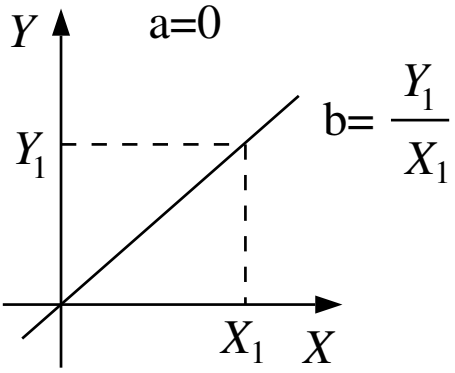
1



2



3



Les constantes b et a sont caractéristiques de la droite :

1. Pour $b > 0$, Y augmente en même temps que X

   Pour $b < 0$, Y diminue quand X augmente

2. Si $b = 0$, la droite est parallèle à l'axe des abscisses

3. Quand $a = 0$, la droite passe par l'origine. Son équation devient $Y = bX$;

   Y et X sont proportionnels, $b = Y/X$ = le coefficient de proportionnalité

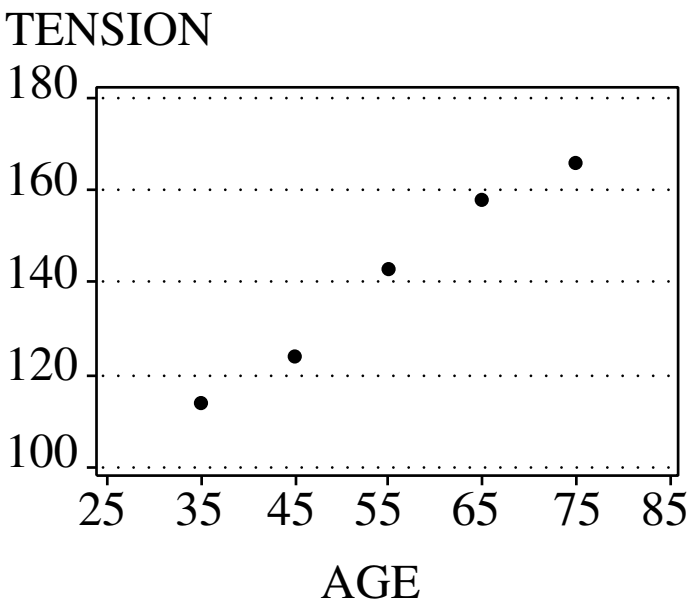
RELATION ENTRE LA TENSION ARTERIELLE ET L'AGE

AGE (X)	TENSION (Y)
35	114
45	124
55	143
65	158
75	166

X = variable explicative = régresseur
Y = variable expliquée

- a et b sont des caractéristiques de la régression. On étudiera comment les évaluer à partir de n couples d'observations (X, Y) non exactement alignés.

- On veut connaître la relation entre la tension artérielle et l'âge. L'étude porte sur 5 femmes âgées respectivement de 35, 45, 55, 65 et 75 ans. Pour chacune, on a mesuré la tension artérielle (en mm de mercure). La tension varie en fonction de l'âge. Existe-t-il une relation ?
- Pour la visualiser, demander aux stagiaires de faire une représentation graphique (distribution de papier millimétré).
- Question : quelle variable en abscisse ?


TENSION
180
160
140
120
100
25
35
45
55
65
75
85
AGE

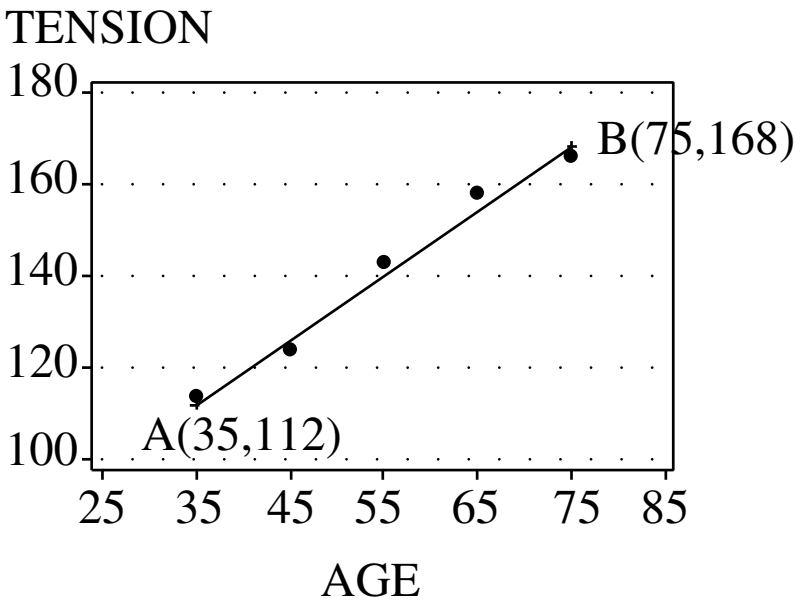
- Présenter le graphique

  X = âge en abscisse

  Y = tension artérielle en ordonnée

  Chaque point représente une femme

- Faire remarquer :
  - la tension augmente avec l'âge
  - cette augmentation semble linéaire
- Faire tracer par les stagiaires la droite qui semble s'ajuster le mieux aux points observés, c'est à dire celle qui passe le plus près possible des points.
- Demander l'évaluation graphique des paramètres a et b.



$$\mathrm{b} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$\mathrm{b} = \frac{168 - 112}{75 - 35}$$

$$\mathrm{b} = 56/40 = 1.4$$

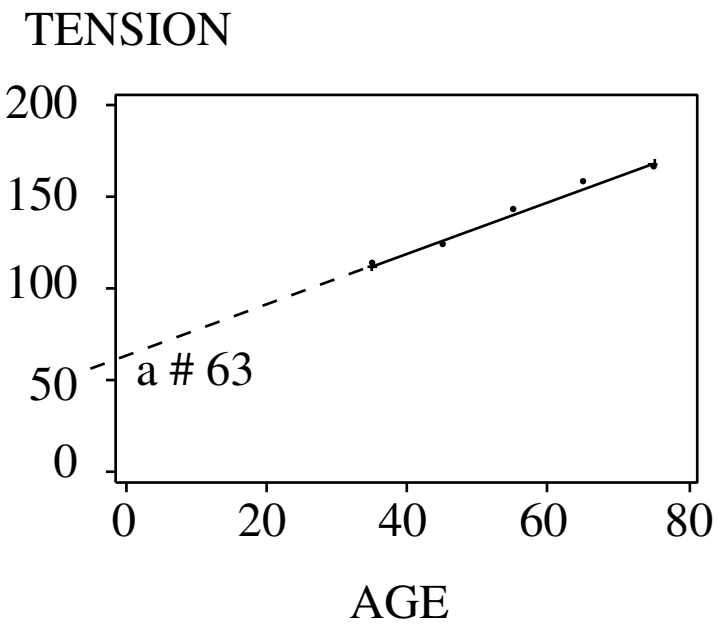
- Présenter l'évaluation graphique de b à partir de la droite Y' tracée par le formateur
  - Soient 2 points A et B sur Y', de coordonnées

$$A(X1 = 35, Y'1 = 112); B(X2 = 75, Y'2 = 168)$$

  - On rappellera la formule du transparent 3 :
    b = différence des ordonnées / différence des abscisses

$$b = (168 - 112)/(75 - 35) = 56/40 = 1.4$$

- La tension augmente de 1.4 mm de mercure par an.



$$Y = \mathrm{a} + \mathrm{b}X$$
$$168 = \mathrm{a} + 1.4 \times 75$$
$$\mathrm{a} = 63$$

$$Y' = 63 + 1.4 \times X$$

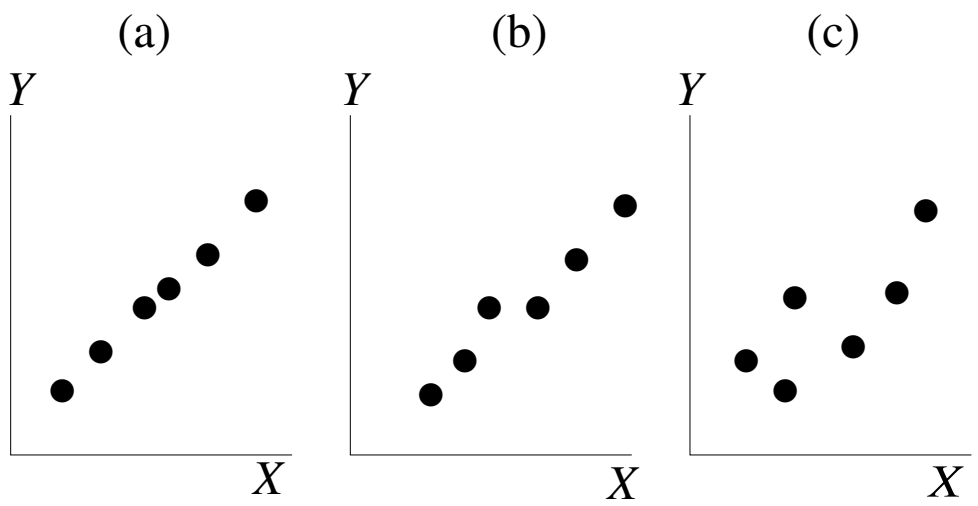
Evaluation graphique de a :

- Constatation : la droite ne passe pas par l'origine. a semble compris entre 60 et 70.
- On peut l'obtenir plus simplement par $Y2 = a + bX2$;

  $168 = a + 1.4 \times 75$

  donc $a = 63$ mm de mercure. Cette valeur devrait représenter la tension à la naissance (âge = 0).
- Conclusion hasardeuse car rien ne permet de penser qu'il existe une relation linéaire entre la tension artérielle et l'âge de 0 à 30 ans. De même, on ne peut dire qu'à 95 ans, la tension sera de 196 mm de mercure.

  Il est déconseillé d'extrapoler hors du domaine observé.



Différents degrés de dispersion $\rightarrow$ Nécessité d'une méthode objective

On recherche une droite dont l'équation est

$$\alpha + \beta X$$

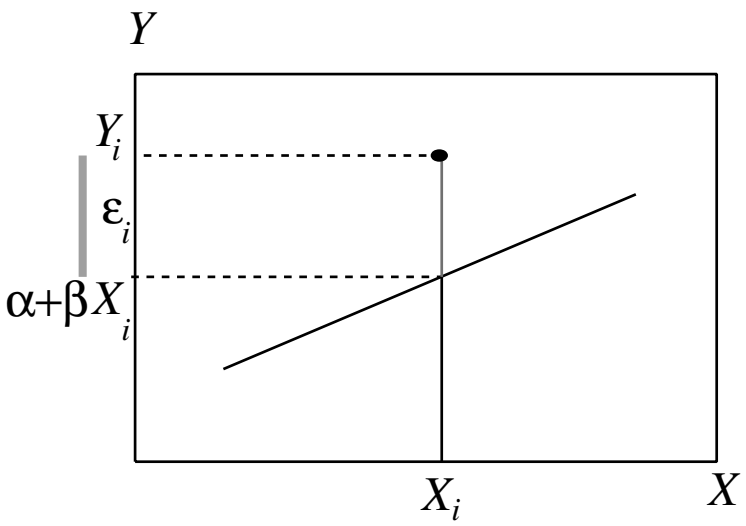
Pour la $i$ème observation, on a : $\alpha + \beta X_i$

$$Y_i - (\alpha + \beta X_i) = \varepsilon_i$$

$$\downarrow \qquad\qquad \downarrow$$

Valeur observée Erreur

- Comparer les couples (a, b) des stagiaires et du formateur.
- Ils doivent être peu différents car les points sont presque alignés. Quel couple (a, b) choisir ?
- L'ajustement est parfait lorsque les points sont parfaitement alignés (fig. a), mais lorsqu'ils sont très dispersés (fig. c), il est difficile de trouver graphiquement la meilleure droite, d'où la nécessité d'une méthode objective qui se prête au calcul.
- Questions :
  - Comment obtenir la droite la "meilleure", c'est à dire celle qui passe le plus près possible des points observés ?
  - Quel critère prendre, c'est-à-dire comment traduire en termes mathématiques cette notion de "meilleure" ?
- Soit : $\alpha + \beta X$ l'équation de la droite que l'on recherche (α et β inconnus).
- Pour une valeur donnée $X_i \Rightarrow \alpha + \beta X_i$ sera différente de Y_i observé.
- On pose $Y_i - (\alpha + \beta X_i) = \varepsilon_i$ avec $\varepsilon_i =$ erreur.
- Si les points étaient parfaitement alignés, on aurait $\varepsilon_i = 0 \; \forall i$



$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Signification graphique de ε_i
  - point observé de coordonnées (X_i, Y_i)
  - $\alpha + \beta X_i$ ordonnée du point d'abscisse X_i situé sur la droite
- On en déduit $Y_i = \alpha + \beta X_i + \varepsilon_i$ (modèle linéaire).
- ε_i = écart parallèle à l'axe des ordonnées.
- Si la question est posée : on ne prend pas l'écart perpendiculairement à la droite, car il ne correspondrait pas au X_i fixé.

Trouver α et β tels que :

1. $\Sigma\, \varepsilon_i$ minimum ?

   mais $\Sigma\, \varepsilon_i = 0$

2. $\Sigma\, \varepsilon_i^2$ minimum

   $\Sigma\, \varepsilon_i^2 = \Sigma\, (Y_i - (\alpha + \beta X_i))^2 = f\,(\alpha, \beta)$

   Critère des moindres carrés (M.C.)

- Comment trouver les estimateurs de α et β ?

  Plusieurs solutions envisageables ; on en donnera 2 :

  1. On choisit α et β de façon à minimiser l'ensemble des erreurs ($\Sigma\,\varepsilon_i$ minimum)
     Quand les points sont parfaitement alignés chaque $\varepsilon_i = 0$. Cette solution ne convient pas car $\Sigma\,\varepsilon_i = 0$. En effet, il y a compensation entre les ε_i situés au- dessus de la droite et ceux en-dessous.
  2. La solution habituelle consiste à prendre le carré de chaque erreur, puis à minimiser la somme : c'est le critère des moindres carrés (M.C.). Il conduit à une droite unique dite droite de régression linéaire (sous-entendu régression obtenue par la méthode des moindres carrés).

$$\Sigma \left(Y_i - (\alpha + \beta X_i)\right)^2 = f(\alpha, \beta) \quad \text{minimum}$$

$$\widehat{\beta} = \frac{\Sigma \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\Sigma \left(X_i - \overline{X}\right)^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\, \overline{X}$$

$\overline{X}$ moyenne des X_i

$\overline{Y}$ moyenne des Y_i

- Dire uniquement que :

  la condition nécessaire pour obtenir le minimum d'une fonction $f(\alpha, \beta)$ est que les dérivées partielles par rapport à α et β soient nulles (l'étude des variations des fonctions repose sur le signe des dérivés d'ordre 1 et 2). On ne détaillera pas.

- On déduit les estimations classiques qui permettent l'estimation de α et β avec les données expérimentales (coefficients de régression)

  β : pente (slope) = accroissement de Y correspondant à l'accroissement d'une unité de X

  α : ordonnée à l'origine (intercept)

Droite *vraie* :

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

Droite *estimée* :

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{X} \implies \widehat{Y} = \overline{Y} + \widehat{\beta}\,(X - \overline{X})$$

Pour la $i$ème observation :

$$\widehat{Y_i} = \overline{Y} + \beta\,(X_i - \overline{X})$$

$\downarrow$

Valeur prédite

$$\text{Pour} \quad X = \overline{X} \quad \implies \quad Y = \overline{Y}$$

La droite passe par le point de coordonnées $(\overline{X}, \overline{Y})$

- Présenter les diverses écritures de la droite de régression estimée :

  la 2e expression s'obtient en remplaçant α par $\overline{Y} - \beta\overline{X}$.

- Montrer que si $X = \overline{X}$, alors $Y = \overline{Y}$

  la droite estimée passe par le point de coordonnées $(\overline{X}, \overline{Y})$;

  $\overline{X}$ = moyenne arithmétique de X;

  $\overline{Y}$ = moyenne de Y.

- Pour la $i$ème observation, $\widehat{Y_i}$ est dite valeur prédite ou ajustée.

X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
35	114	-20	-27	400	540
45	124	-10	-17	100	170
55	143	0	2	0	0
65	158	10	17	100	170
75	166	20	25	400	500
275	705	0	0	1000	1380

$$\overline{X} = 275/5 = 55 \qquad \overline{Y} = 705/5 = 141$$

$$\widehat{\beta} = 1380/1000 = 1.38$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{X} = 141 - 1.38 \times 55 = 65.1$$

- Estimation de α et β sur l'exemple tension/âge par la méthode M.C. (soit par les stagiaires, soit en présentant chaque colonne successivement)

1. Calcul des moyennes $\overline{X}$ et $\overline{Y}$
2. Calcul, pour chaque observation, des écarts $(X_i - \overline{X}), (Y_i - \overline{Y})$.
   Faire remarquer $\Sigma(X_i - \overline{X}) = 0$ et $\Sigma(Y_i - \overline{Y}) = 0$
3. Calcul des carrés des écarts $(X_i - \overline{X})^2$ et de leur somme
4. Calcul des produits $(X_i - \overline{X})(Y_i - \overline{Y})$ et de la somme
5. Calcul de $\hat{\alpha}$ et $\hat{\beta}$ et montrer que les valeurs sont différentes de a et b

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}\, X = 65.1 + 1.38 \times X$$

Valeur prédite pour $X = 35$

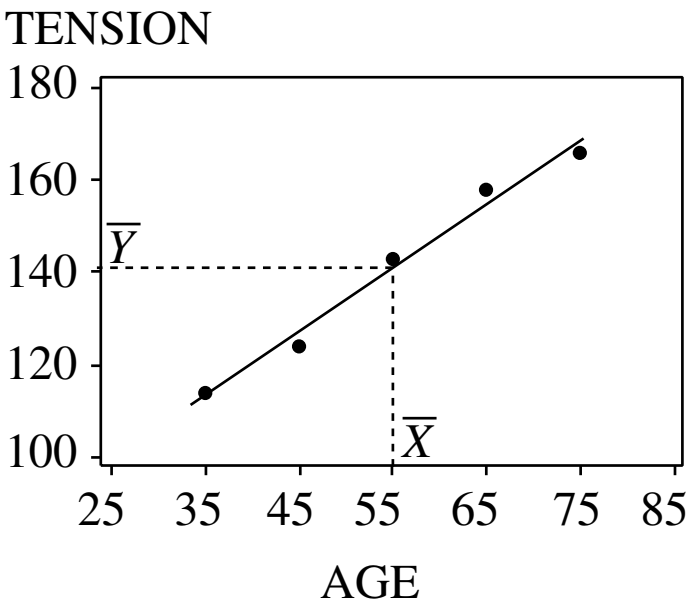
$$\widehat{Y} = 65.1 + 1.38 \times 35 = 113.4$$

Valeur prédite pour $X = 55 = \overline{X}$

$$\widehat{Y} = 141 = \overline{Y}$$

- Connaissant $\hat{\alpha} = 6.51$ et $\hat{\beta} = 1.38$, on peut tracer la droite estimée à partir des coordonnées de 2 points :
  - 1er point : valeur prédite de Y pour $X = 35 \Rightarrow \widehat{Y} = 113.4$
  - 2e point : $X = 55 = \overline{X} \Rightarrow \widehat{Y} = 141 = \overline{Y}$; on vérifie ainsi que la droite passe bien par le point de coordonnées égales aux moyennes $(\overline{X}, \overline{Y})$

$$\widehat{Y} = 65.1 + 1.38X$$



Présenter le graphique des points observés et de la droite prédite ou ajustée.

X_i	Y_i	$\widehat{Y_i}$	$\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$	$\widehat{\varepsilon_i}^2 = (Y_i - \widehat{Y_i})^2$
35	114	113.4	0.6	0.36
45	124	127.2	-3.2	10.24
55	143	141.0	2.0	4.00
65	158	154.8	3.2	10.24
75	166	168.6	-2.6	6.76
	705	705	0	31.60

$$\sum \widehat{\varepsilon_i} = \sum \left(Y_i - \widehat{Y_i}\right) = 0$$

$$\sum \widehat{\varepsilon_i}^2 = \sum \left(Y_i - \widehat{Y_i}\right)^2 = 31.60$$

$$\text{Carré moyen résiduel} = \frac{\sum \widehat{\varepsilon_i}^2}{n-2}$$

$$CMR = \frac{31.60}{3} = 10.533$$

- Calcul de $\sum \varepsilon_i^2$
  - On commence par calculer les valeurs prédites $\widehat{Y}_i$ pour chaque observation
    (faire remarquer $\sum Y_i = \sum \widehat{Y}_i$)
  - Calcul du résidu correspondant $\hat{\varepsilon}_i$
    (faire remarquer $\sum \hat{\varepsilon}_i = 0$)
  - Porter chaque résidu au carré et faire la somme
    $\sum \hat{\varepsilon}_i^2 = 31.60$
- Signaler que l'on appelle CARRE MOYEN RESIDUEL la quantité $\sum \hat{\varepsilon}_i^2/(n-2)$ soit C.M.R. $= 31.60/3 = 10.53$;
  $(n-2)$ = nombres de degrés de liberté de la variabilité résiduelle ; son calcul sera expliqué ultérieurement.
- On trouve également l'expression CARRE MOYEN DE L'ERREUR.
- Résiduel = ce qui reste quand on a tenu compte de la régression, c'est à dire ce qui n'est pas expliqué par la régression.

Calcul à partir de a et b évalués graphiquement

X_i	Y_i	$Y_i^{'}$	$Y_i - Y_i^{'}$	$(Y_i - Y_i^{'})^2$
35	114	112	2	4
45	124	126	-2	4
55	143	140	3	9
65	158	154	4	16
75	166	168	-2	4
			5	37

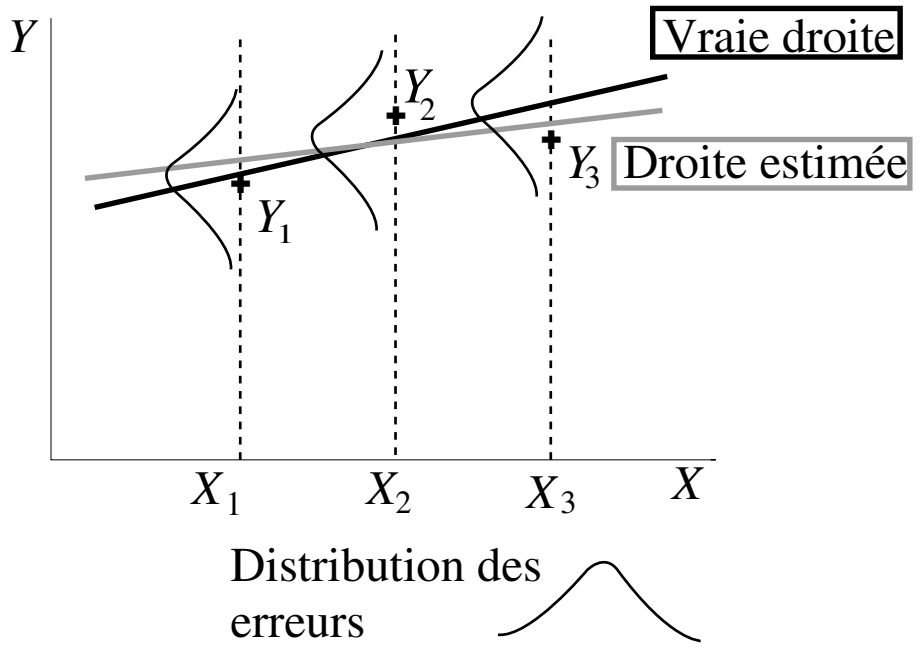
$$Y_i^{'} = \mathrm{a} + \mathrm{b}\ X_i = 63 + 1.4 \times X_i$$

$$\underset{\text{moindres carrés}}{31.60} \quad < \quad \underset{\text{méthode graphique}}{37}$$

(facultatif)

Permet de vérifier que $\Sigma\,\hat{\varepsilon}_i^2$ par M.C. est inférieure à celle obtenue par la méthode graphique ($a = 63$ et $b = 1.4$).

On appellera Y' la valeur prédite par méthode graphique.



$$Y = \alpha + \beta X \quad \text{inconnue}$$

$$Y_1, Y_2, Y_3 \text{ mesures en } X_1, X_2, X_3$$

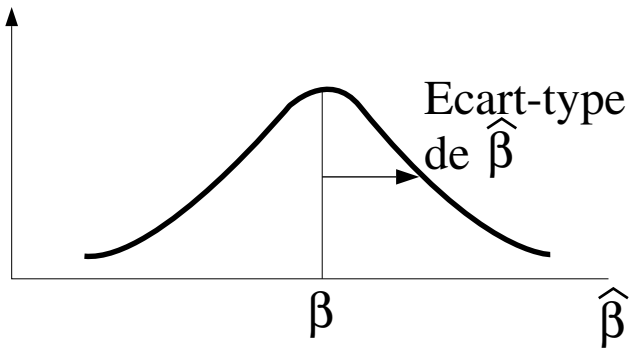
$$\Longrightarrow \widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$$

Si on refait des expériences identiques $\Rightarrow$ d'autres valeurs $\widehat{\alpha}$, $\widehat{\beta}$

- Par la méthode des moindres carrés, on a obtenu une droite estimée $\widehat{\alpha} + \widehat{\beta}X$,

  la vraie droite $\alpha + \beta X$ étant inconnue.

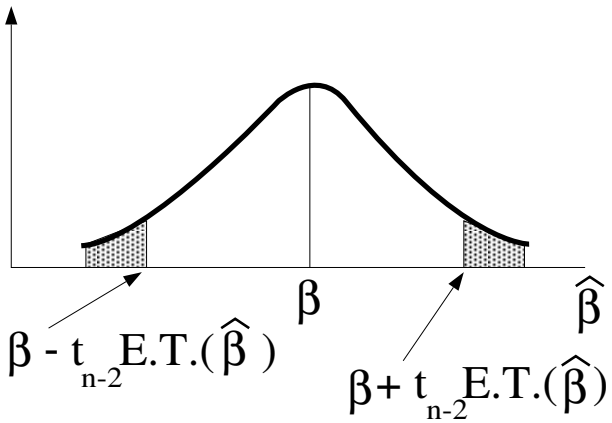
- Si on recommence l'expérience dans les mêmes conditions, on trouvera des valeurs de $\widehat{\alpha}$ et de $\widehat{\beta}$ différentes. On en déduit que $\widehat{\alpha}$ et $\widehat{\beta}$ sont des valeurs prises par des variables aléatoires et, qu'à ce titre, elles ont une espérance et une variance.



$$\text{Ecart} - \text{type de } \widehat{\beta} = \frac{\sigma}{\sqrt{\Sigma \left(X_i - \overline{X}\right)^2}}$$

Estimation de σ^2 par :

$$CMR = \frac{\Sigma \left(Y_i - \widehat{Y_i}\right)^2}{n-2} = \frac{\Sigma \, \widehat{\varepsilon_i}^2}{n-2}$$



- A partir de diverses suppositions sur les erreurs ε_i, qui seront développées ultérieurement, on montre que l'estimateur $\widehat{\beta}$ est distribué approximativement selon une loi gaussienne avec comme espérance β et comme écart-type

$$E.T.(\widehat{\beta}) = \sigma / \sqrt{\sum (X - \overline{X})^2}$$

- σ représente l'écart-type des observations Y autour de la vraie droite ;

  σ^2 est inconnu et on l'estime par le carré moyen résiduel (écarts résiduels entre Y et la droite ajustée).

- L'utilisation de C.M.R. pour estimer σ^2 introduit une source supplémentaire de non fiabilité, particulièrement si n est petit, ce qui conduit à remplacer la distribution normale par une distribution similaire, celle de STUDENT à $(n-2)$ d.d.l.

- Montrer sur la distribution que l'on peut déduire une plage de valeurs possibles pour β inconnu constituant ainsi un intervalle de confiance.

Intervalle de confiance (I.C.) de la pente :

$$\widehat{\beta} \pm \mathrm{t_{n-2}} \times \mathrm{E.T.}\left(\widehat{\beta}\right)$$

$$\text{variance}\left(\widehat{\beta}\right) = \frac{CMR}{\Sigma\left(X_i - \overline{X}\right)^2} = \frac{10.533}{1000} = 0.0105$$

$$\mathrm{E.T.}\left(\widehat{\beta}\right) = \sqrt{0.0105} = 0.1026$$

Pour un niveau de confiance $= 0.95$ $\mathrm{t_{n-2}} = 3.182$

I.C. :

$1.38 \pm 3.182 \times 0.1026$

limite supérieure 1.706

limite inférieure 1.050

$\Big\}\ 1.05 < \beta < 1.71$

L'intervalle de confiance ne contient pas 0

- Calculer l'estimation de β par intervalle de confiance : sur l'exemple tension/âge (analogie avec intervalle de confiance de la moyenne)
  - écart-type de $\widehat{\beta} = 0.1026$
  - les limites de l'intervalle pour une confiance de 0.95 sont 1.71 et 1.05. On constate que l'intervalle ne contient pas 0 ; cette valeur est donc peu vraisemblable et β doit certainement différer de 0.
- Rappel : $\beta = 0$ signifie que Y ne varie pas quand X varie ; la droite est parallèle à l'axe des abscisses : $\widehat{Y} = \overline{Y}$.

Test de l'hypothèse nulle $H_0 : \beta = 0$

Si H_O est vraie on montre que :

$$T = \frac{\widehat{\beta}}{\text{E.T.}\left(\widehat{\beta}\right)} \rightsquigarrow \text{distribution de Student à } n-2 \text{ d.d.l.}$$

$$\text{Pour} \quad \widehat{\beta} = 1.38 \qquad \text{E.T.}\left(\widehat{\beta}\right) = 0.1026$$
$$T = 1.38/0.1026 = 13.45$$

Risque	t . Student	
0.05	3.182	$T >$ t
0.01	5.841	$T >$ t

Rejet de $H_0 \Rightarrow$ rejet de $\beta = 0$

- On peut considérer un intervalle de confiance comme l'ensemble des hypothèses acceptables pour β.
- Avec les tests d'hypothèses on se concentre sur une seule hypothèse dite hypothèse nulle H_0.
- Celle qui nous intéresse ici :

  $$H_0 : \beta = 0$$

  (la tension n'augmente pas avec l'âge)

  on va rechercher dans quelle mesure elle est compatible avec les données.
- On montre que si H_0 est vraie ($\beta = 0$), la statistique $T = \widehat{\beta}/E.T.(\widehat{\beta})$ est une variable de STUDENT à $(n-2)$ degrés de liberté.
- Application :

  $\widehat{\beta} = 1.38, E.T.(\widehat{\beta}) = 0.1026$ entraîne $T = 13.45$. On compare T à une valeur seuil trouvée dans la table de STUDENT pour un risque de 0.05, 0.01...
- A l'intersection de la ligne $(n-2) = 3$ et de la colonne 0.05, on trouve la valeur 3.182, $T > 3.182$, $T > 5.84$ donc $P < 0.01$. En conclusion, l'hypothèse nulle est peu crédible et nous pouvons la rejeter.

Intervalle de confiance (I.C.) de α :

$$\widehat{\alpha} \pm \mathrm{t_{n-2}} \times \mathrm{E.T.}\left(\widehat{\alpha}\right)$$
$$\left(\overline{Y} - \widehat{\beta}\,\overline{X}\right) \pm \mathrm{t_{n-2}} \times \mathrm{E.T.}\left(\overline{Y} - \widehat{\beta}\,\overline{X}\right)$$

$$\mathrm{variance}\left(\widehat{\alpha}\right) = CMR\left(\frac{1}{n} + \frac{\overline{X}^2}{\Sigma\left(X_i - \overline{X}\right)^2}\right)$$

$$\mathrm{E.T.}\left(\widehat{\alpha}\right) = \sqrt{\mathrm{variance}\left(\widehat{\alpha}\right)}$$

I.C. :
$65.1 \pm 3.182 \times 5.828$
$46.6 < \alpha < 83.6$

facultatifs (transparents 23 et 24)

- On procède de façon analogue pour le calcul de l'intervalle de confiance de α et le test de $H_0 : \alpha = 0$ (moins intéressant que pour β).
- Rappel : α n'a de sens que si l'on accepte la linéarité à partir de $X = 0$.
- La variance de α contient un terme supplémentaire car $\alpha = Y - \beta X$.
- L'intervalle de confiance de α pour l'exemple ne contient pas 0.

Test de l'hypothèse nulle $H_0 : \alpha = 0$

$$T = \frac{\widehat{\alpha}}{\text{E.T.}\,(\widehat{\alpha})}$$

$$T = 11.17$$

(facultatif)

- $H_0 : \alpha = 0$.
- Demander la valeur seuil servant à la comparaison.
- La relation ne passe pas par l'origine (supposition : linéarité de 0 à 75 ans).

$$\Sigma\,\varepsilon_i^2 = \Sigma\left(Y_i - \widehat{Y_i}\right)^2 = 31.60$$
$$= \text{ variabilité résiduelle}$$

$$\Sigma\left(Y_i - \overline{Y}\right)^2 = (114-141)^2 + (124-141)^2 + \cdots + (166-141)^2$$

$$\Sigma\left(Y_i - \overline{Y}\right)^2 = 1936 = \text{variabilité totale}$$

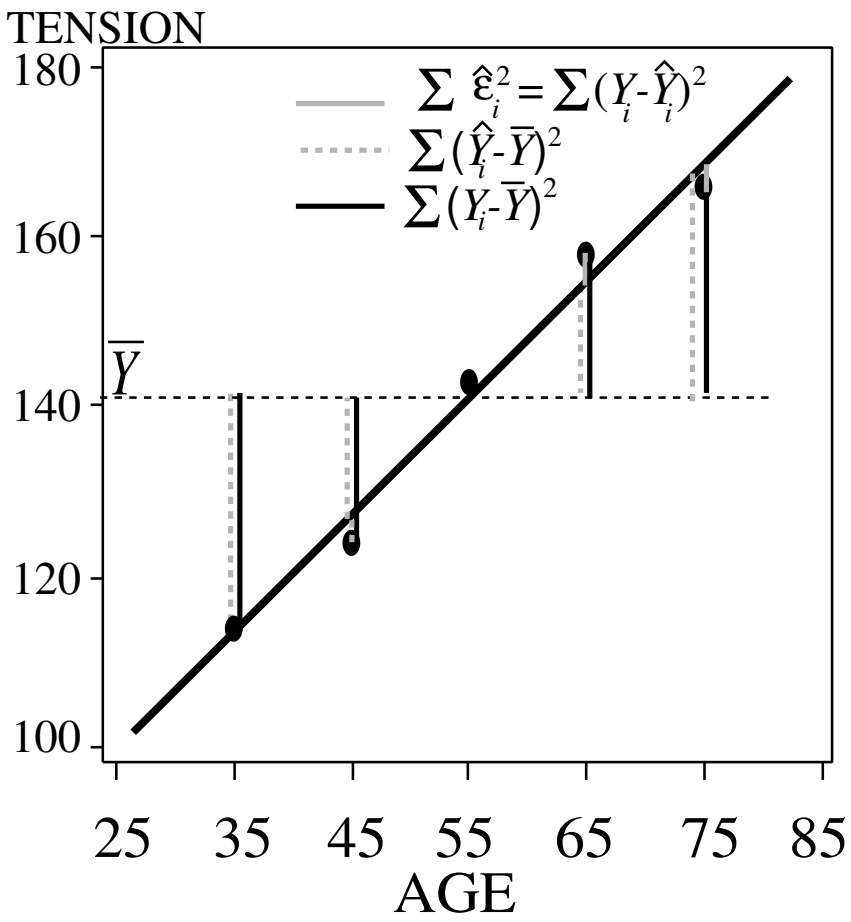
Que représente : $1936 - 31.60 = 1904.4$?

$\rightarrow$ ce qui est expliqué par la régression

$$\Longrightarrow \Sigma\left(\widehat{Y_i} - \overline{Y}\right)^2$$

- La méthode des M.C. est largement utilisée car :
  1. elle conduit à des formules simples pour estimer α et β ;
  2. elle est étroitement liée à l'analyse de variance qui sera étudiée ultérieurement.
- Nous allons montrer que l'on peut exprimer les résultats de la régression sous la forme d'analyse de variance.
- Rappel : la variabilité résiduelle $\sum \hat{\varepsilon}_i^2 = 31.60$ représente la somme des carrés des écarts entre les valeurs Y observées et Y prédites.
- Montrer que si l'on ne tient pas compte de l'information apportée par la relation entre Y et X, on peut calculer une autre variabilité que l'on appelle variabilité totale. On l'obtient en mettant au carré la différence entre les valeurs observées de Y et la moyenne $\overline{Y}$ et en effectuant la somme, soit 1936.
- Quelle est la signification de la différence $1904.4 = 1936 - 31.60$?

  Elle correspond à la somme des carrés des écarts des valeurs prédites à la moyenne globale. On l'appelle somme des carrés expliquée par la régression (ou par le modèle).

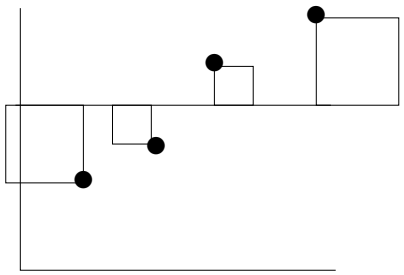


$$\widehat{Y} = 65.1 + 1.38 \times X$$

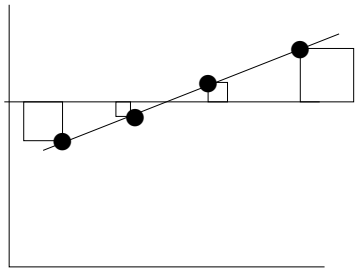
(transparents 26 ou 26bis et 27 au choix)

Schématisation graphique des 3 types de variabilité qui conduit à l'analyse de variance.

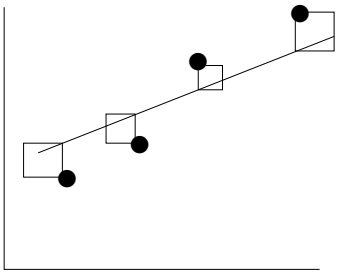
Carrés



totaux



dus à la régression



résiduels

(transparents 26 ou 26bis et 27 au choix)

Schématisation graphique des 3 types de variabilité qui conduit à l'analyse de variance.

VARIABILITE TOTALE
$= 1936 = \sum (Y_i - \bar{Y})^2$

ERREUR
$= 31.60 = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2$

REGRESSION
$= 1904.4 = \sum (\hat{Y}_i - \bar{Y})^2$

(transparents 26 ou 26bis et 27 au choix)

Equation d'analyse de la variance :

$$\Sigma (Y_i - \overline{Y})^2 = \Sigma \left(\widehat{Y}_i - \overline{Y}\right)^2 + \Sigma \left(Y_i - \widehat{Y}_i\right)^2$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

variation totale	variation expliquée par la régression	variation résiduelle

Nombre de degrés de liberté :

$$n - 1 = 1 + (n - 2)$$

- Présenter l'équation d'analyse de la variance :

  variation totale = variation expliquée par la régression + variation résiduelle

- A chaque somme des carrés correspond des degrés de liberté (d.d.l.) :

  $(n-1)$ = nombre de relations indépendantes. Il y a n écarts $(Y_i - \overline{Y})$ mais, comme $\Sigma(Y_i - \overline{Y}) = 0$, le dernier est déterminé à partir des $(n-1)$ premiers; il n'y a donc que (n - 1) écarts indépendants.

  1 = d.d.l régression. Il y a n écarts $(\widehat{Y_i} - \overline{Y})$; il suffit de connaître un seul $\widehat{Y_i}$ pour en déduire tous les autres, il y a donc une seule relation indépendante.

  $(n-2)$ = d.d.l. résiduel. On l'obtient par différence entre d.d.l. total et d.d.l. de la régression, soit :

  $$(n-1) - 1 = n - 2$$

  ce qui donne, pour l'exemple tension/âge :

  d.d.l. total = 5 - 1 = 4

  d.d.l. régression = 1

  d.d.l. résiduel = n - 2 = 5 - 2 = 3

Source de variation	Somme des carrés	DDL	Carré moyen
expliquée par la régression (modèle)	$\Sigma\left(\widehat{Y_i}-\overline{Y}\right)^2$ 1904.4	1	CMM 1904.4
due aux erreurs	$\Sigma\left(Y_i-\widehat{Y_i}\right)^2$ 31.6	$n-2$ 3	CMR 10.53
totale	$\Sigma\left(Y_i-\overline{Y}\right)^2$ 1936	$n-1$ 4	

$$F=\frac{CMM}{CMR}=\frac{1904.4}{10.53}=180.8$$

- A partir de l'équation d'analyse de la variance, construction du tableau d'analyse de la variance. Le commenter avec l'exemple, dire que les principaux logiciels fournissent les résultats sous cette forme (sans les formules).
- Introduire le test de la signification de la régression par le F de FISHER- SNEDECOR.
- Si H_0 est vraie ($\beta = 0$), on montre que le rapport CM régression/CMR est une variable de FISHER-SNEDECOR (table à 1, $n - 2$ d.d.l.).
- Si F observé $>$ F table pour un risque donné, on rejette H_0.
- Recherche du F dans la table pour l'exemple tension/âge.

Mesure de l'ajustement

Coefficient de détermination :

$$R^2 = \frac{\Sigma\left(\widehat{Y_i} - \overline{Y}\right)^2}{\Sigma\left(Y_i - \overline{Y}\right)^2}$$

$$\frac{\text{VARIATION} \quad \text{expliquée par le modèle}}{\text{VARIATION} \quad \text{totale}}$$

$$\text{ex :} \quad R^2 = \frac{1904.4}{1936} = 0.98$$

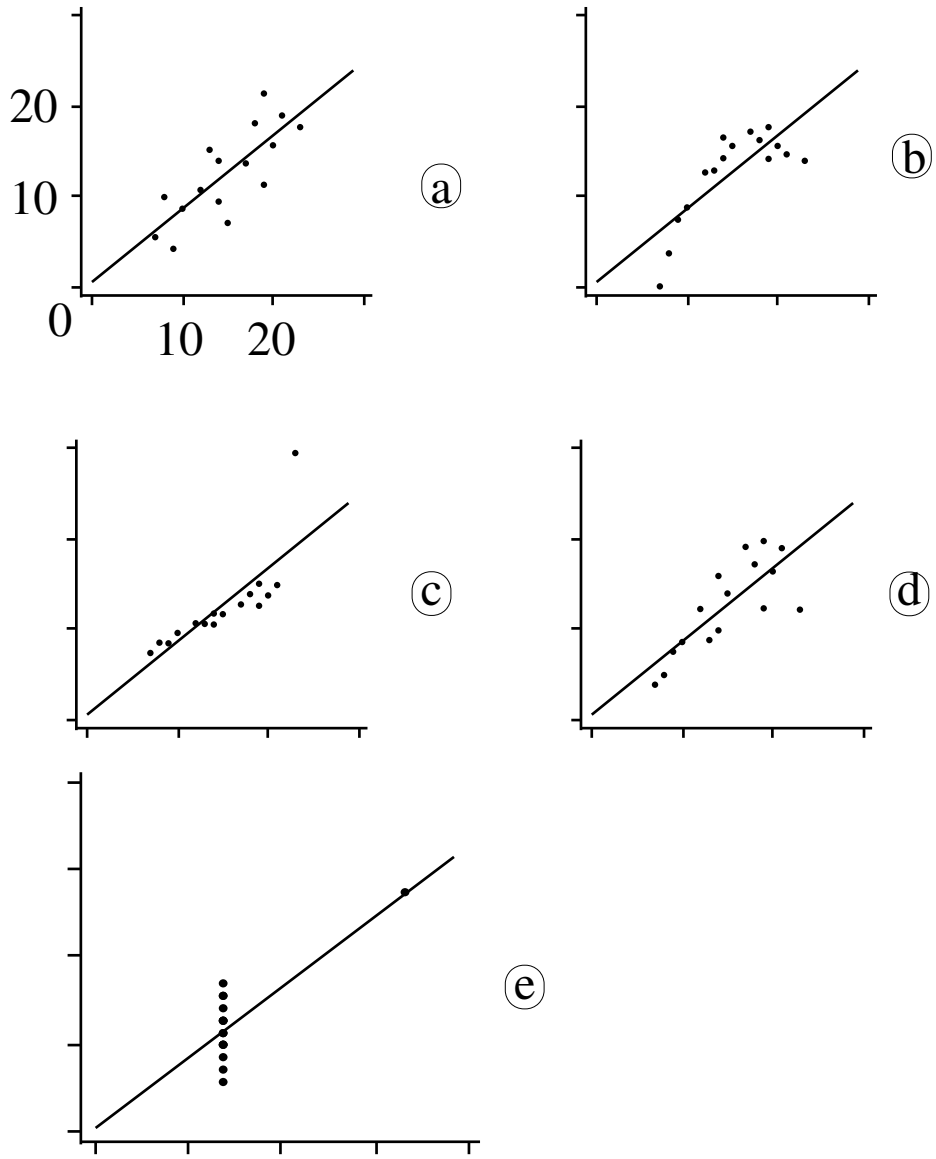
La régression explique 98% de la variabilité totale

- Introduire un indice couramment employé pour donner une valeur à la qualité de l'ajustement du modèle : il s'agit du coefficient de détermination R^2.
- R^2 = variabilité expliquée par le modèle divisée par la variabilité totale.
- R^2 est la proportion de la variabilité totale de Y expliquée par$X$. Si la valeur de R^2 est proche de 1, on dira que X explique bien Y : $R^2 = 0.98$ l'âge explique l'augmentation de tension

  (facultatif) Signaler qu'on peut utiliser une statistique de test pour savoir si R^2 est significative :

  $$t = R\sqrt{n-2}/\sqrt{1-R^2}$$

  où t est la valeur absolue d'une variable de STUDENT à $(n-2)$ d.d.l.


20
10
0
10
20
a
b
c
d
e

- Evaluation critique de l'ajustement.
  - De quels moyens dispose-t-on pour affirmer que les résultats issus de l'analyse de régression (obtenus aisément grâce aux logiciels stat.) sont en accord avec les données expérimentales?
  - Les méthodes graphiques sont parmi les plus simples et elles permettent une vérification valable ; elles devraient être effectuées avant tous calculs statistiques.
- Donner l'exemple des 5 ensembles de données de 16 couples (X, Y).
  - Un programme de régression standard fournit pour chaque ensemble les mêmes statistiques, à savoir :

    $\widehat{\beta} = 0.809$, $\widehat{\alpha} = 0.52$

    $R^2 = 0.617$
  - Les 5 ensembles présentent la même droite de régression, cependant si on examine les graphiques, on s'aperçoit que le modèle linéaire ne s'adapte pas à tous les cas. En particulier:

    fig. b : les points se répartissent selon un arc, ce qui laisse penser que la liaison est de type quadratique (terme en X^2) ;

    fig. c : on note la présence d'un point aberrant. Il conviendrait de vérifier les données. Y a-t-il une cause qui justifie cet écart ?

    fig. d : dispersion des Y augmente en même temps que X. Les suppositions sur les ε_i ne sont pas respectées. Une transformation, type *log*, sur les Y peut être envisagée ;

    fig. e : 1 point très influent. Mauvais plan d'expérience.