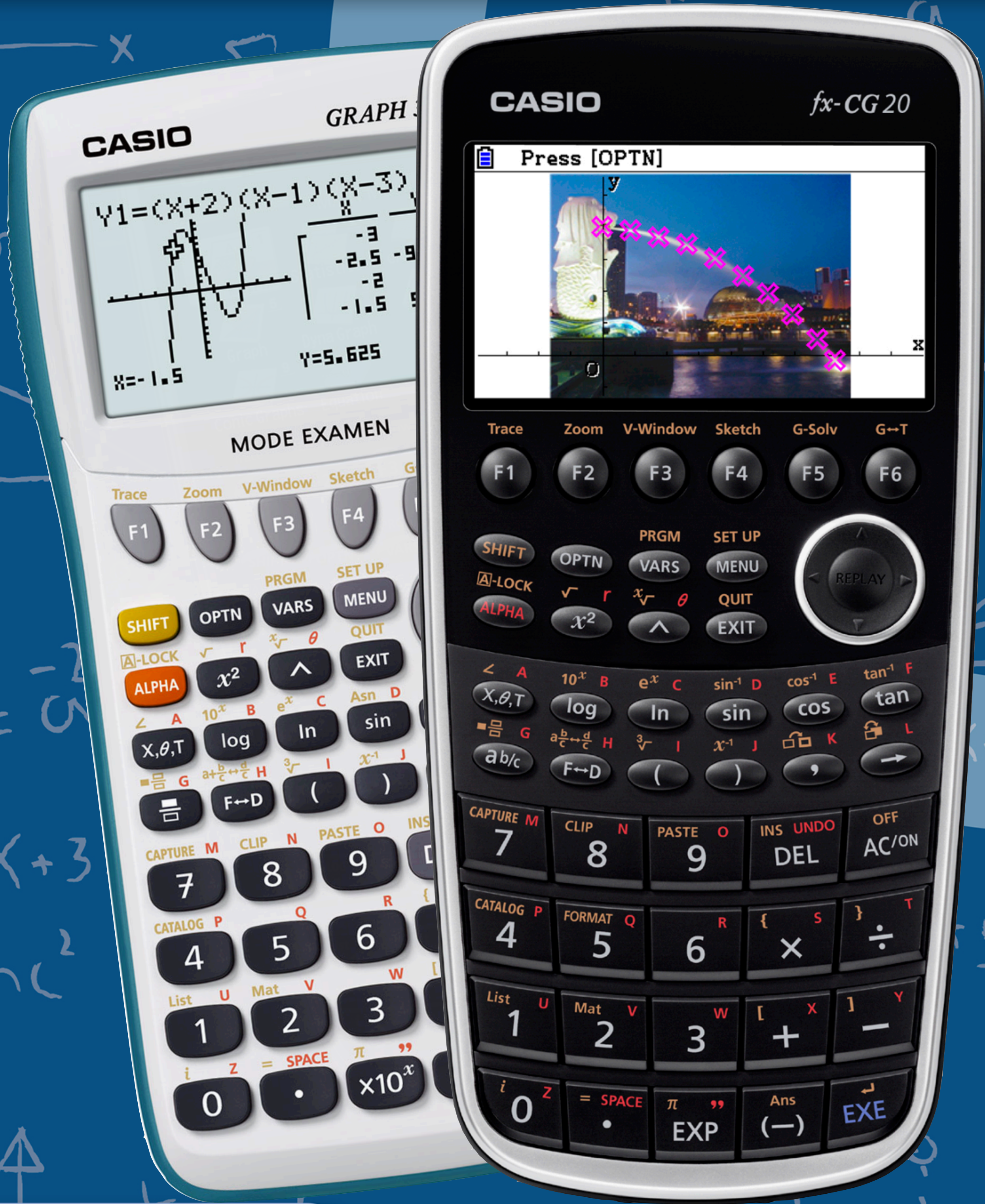
Statistiques à 2 variables
… et calculatrice graphique

Congrès SBPMef Mons 2015



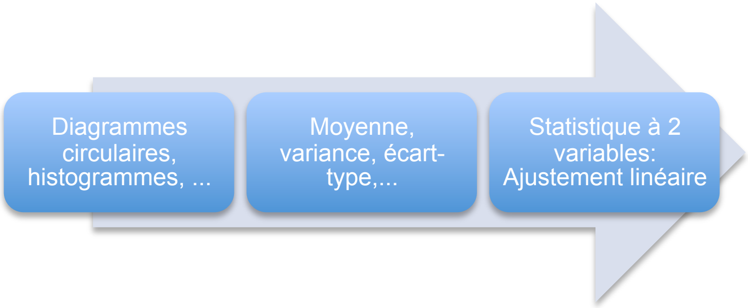
www.casio-education.be

CASIO®

Merci,
L'équipe CASIO Education

La droite de régression :

construction intuitive des paramètres
de la 25+ Pro à la Fx-CG20



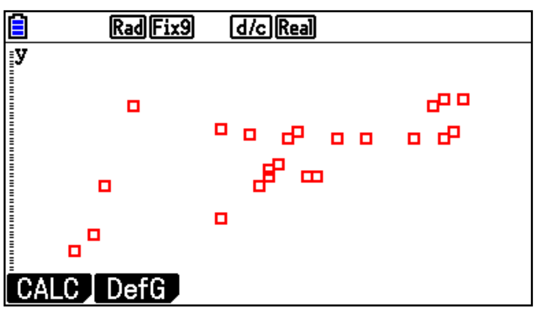
1. Qu'est-ce que la régression ?

Dans la vie réelle, il est courant de travailler sur des statistiques à plusieurs variables. Pour une statistique à plusieurs variables, si les données sont quantitatives, nous pouvons déjà appréhender une bonne part de la réalité avec les outils dont nous disposons pour les statistiques à une variable **(moyenne, écart-type, diagramme des fréquences cumulées etc.).**

Mais ensuite, il est intéressant de déterminer s'il existe une relation entre certaines de ces variables. Et si cette relation existe, est-elle forte ou pas ?

Pour des couples de données quantitatives, il est possible de représenter les résultats de ces mesures dans un système d'axes. Pour discerner plus facilement s'il existe une dépendance ou non entre les deux observations, le plus simple est de représenter chaque couple d'observation dans un repère cartésien. On parle alors de nuage de points ou de diagramme de dispersion.

Par exemple, en affichant les tailles et poids d'un groupe de personnes dans un repère, nous observons que les données se répartissent de la manière suivante :

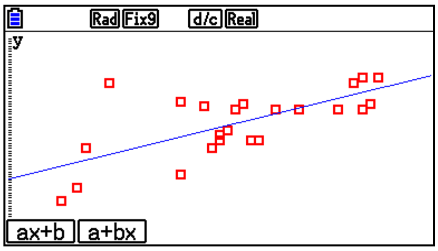


Le poids est-il lié à la taille ?

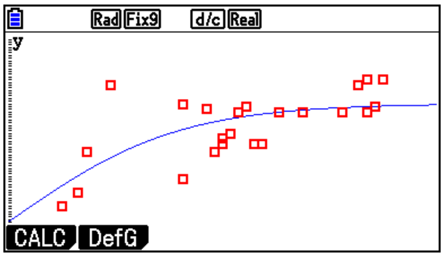
Intuitivement, nous avons souvent tendance à vouloir étudier des phénomènes continus. Donc, s'il existe un lien entre les deux variables étudiées, nous voudrions **ajuster aux observations une courbe d'estimation rendant**

compte au mieux de ce lien, nous donnant ainsi une possibilité d'interpolation et d'extrapolation. C’est le propos de la régression.

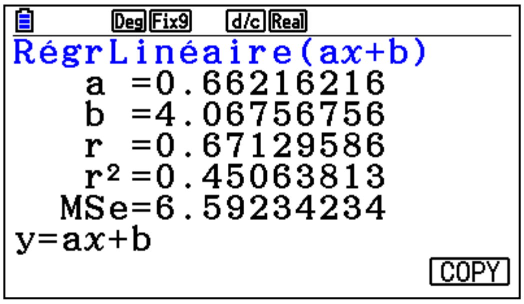
La calculatrice permet de calculer facilement des courbes d’ajustement possibles pour ces données, comme par exemple, la droite de régression, la régression logarithmique ou une fonction polynomiale de degré 4.

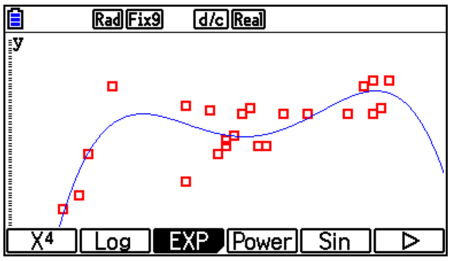


Mais laquelle est la meilleure ?



Pour répondre à cette question, nous avons deux mesures qui sont le coefficient de corrélation et le coefficient de détermination.





Mais que représentent-ils ?

En nous inspirant d’un exercice de Droesbeke [1], nous présentons dans un premier temps la méthode pour calculer la droite de régression avec la calculatrice et interpréter les résultats visuellement.

Ensuite nous reprenons le même exercice pour construire de manière intuitive les paramètres de la droite de régression (pente et ordonnée à l’origine) et les coefficients de corrélation et de détermination qui nous permettront de juger de la qualité de l’ajustement.

[1] *Eléments de Statistique.* DROESBEKE, Editions de l'Université de Bruxelles, Bruxelles 1992

2. Calcul « rapide » de la droite de régression

Intéressons-nous à l'argent de poche donné à des jeunes dont l'âge est compris entre 11 et 16 ans…

Sur une année, on a collecté, pour 10 adolescents l'âge et le montant hebdomadaire moyen, exprimé en euros, en notant le couple (âge, montant) dans le tableau ci-contre.

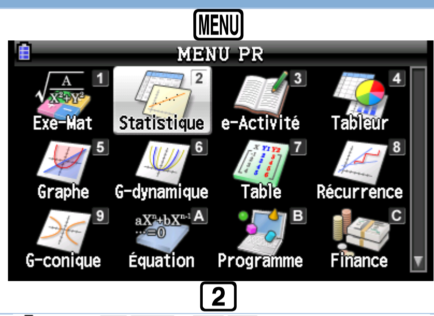
AGE	ARGENT
12	4,1
12	3,4
15	11,3
14	10,2
16	11,5
14	7,2
12	6,0
13	7,8
11	3,5
11	3,0

Nous allons d'abord découvrir avec la calculatrice comment

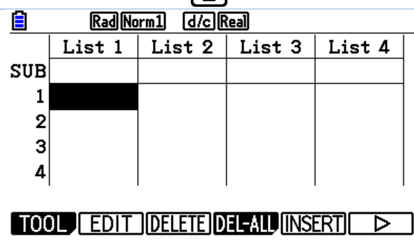
- *afficher le nuage de point ;*
- *calculer le point moyen et le situer par rapport au nuage ;*
- *déterminer les paramètres de la droite de régression et les coefficients de corrélation et détermination ;*
- *superposer point moyen, droite de régression et nuage de points ;*
- *estimer l'argent de poche d'un jeune de 18 ans et d'un nouveau-né.*

1. Entrer les données de l'échantillon

- Pour entrer les données de l'échantillon, sélectionner via MENU l'icône *Statistique*, en tapant directement MENU 2 ou via le curseur en validant avec EXE.



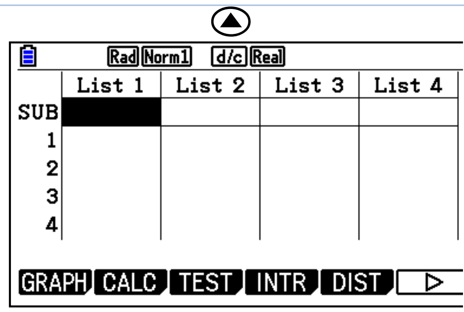
- Un tableau apparaît, composé de 26 colonnes intitulées successivement List1, List2, List3 etc. Elles sont destinées à contenir les observations. Le déplacement dans les lignes et les colonnes se fait au moyen du curseur.



- En dessous de chaque titre de colonne, un emplacement est réservé pour entrer éventuellement le libellé des colonnes

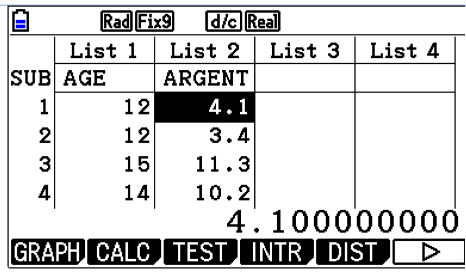
→ Se positionner au moyen du curseur en dessous de List1

→ Si des données sont déjà présentes, réinitialiser la liste entière en se positionnant avec le curseur dans la colonne à supprimer et sélectionner *DEL-A* [F6] [F4].



- Pour entrer le titre de la colonne, taper les lettres correspondantes (affichées en rouge au dessus des touches et sélectionnables au moyen de la touche [ALPHA]).

- Pour entrer les différentes valeurs, se positionner au moyen du curseur à la première ligne de la première colonne et encoder les valeurs, validées chacune par [EXE].

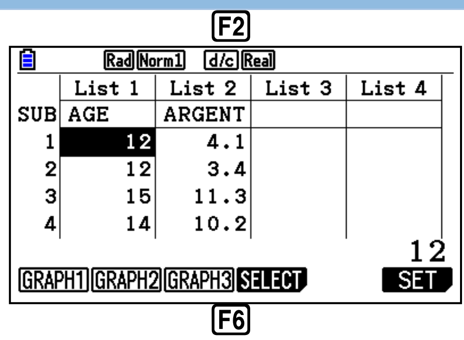


2. AFFICHER LE NUAGE DE POINTS

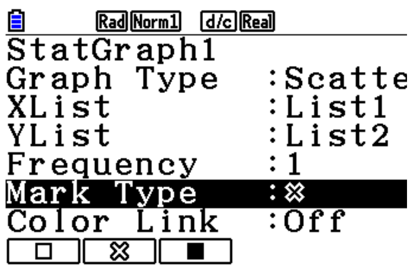
→ S'assurer que l'ajustement automatique de la fenêtre aux données est sélectionné via [SHIFT] [MENU] [F1] (l'item StatWind à Auto).

→ Dans l'écran principal, choisir l'option GRAPH [F1].

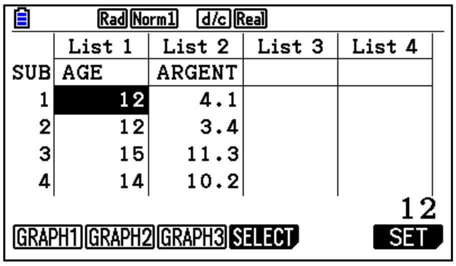
→ Taper [F6] *SET* pour associer les colonnes considérées et le type de graphe à afficher.



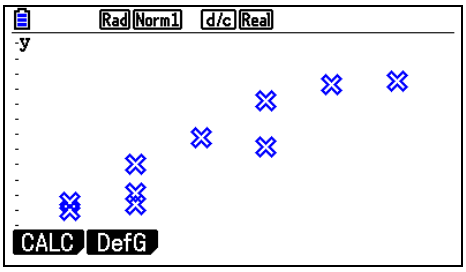
→ Dans notre cas, nous choisissons
  - afficher des points (*Scatter = dispersion, pétale, confetti, etc.)*
  - sélectionner *List1* pour l'âge (X), *List2* pour l'argent (Y) et 1 pour les fréquences puisque les données ne reviennent qu'une fois (si nécessaire, descendre sur ces items au moyen du curseur et entrer le numéro de la liste correspondante en validant avec [EXE]).Une fois les paramètres validés, sortir du menu avec [EXIT].



→ Sélectionner le graphe GRAPH1 F1 et le nuage de point s'affiche.



En observant le nuage de points, nous constatons qu'il est concentré, que même s'il ne s'agit pas d'une droite, il y a une direction globale du nuage de points, une orientation : il semble que les adolescents les plus âgés reçoivent d'avantage d'argent que leurs cadets. Une droite qui traduirait cette orientation serait de pente positive.

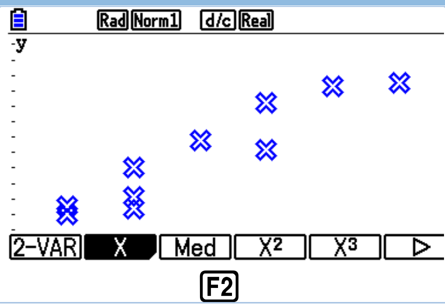


Tenter de modéliser ce problème par une droite de régression semble donc une démarche valide.

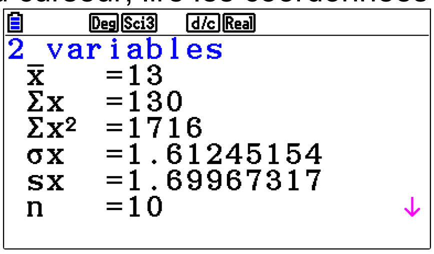
3. Calculer le point moyen et le superposer au nuage de points

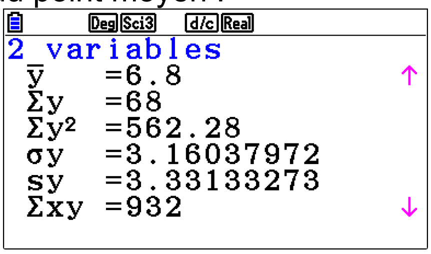
→ Sélectionner F1 Calc.

→ Sélectionner *2Var* F1 pour calculer tous les indices de position et dispersion des 2 variables (les autres items concernent les différentes régressions).



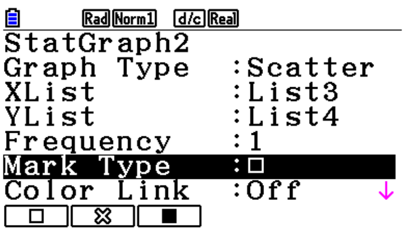
→ Au moyen du curseur, lire les coordonnées du point moyen :

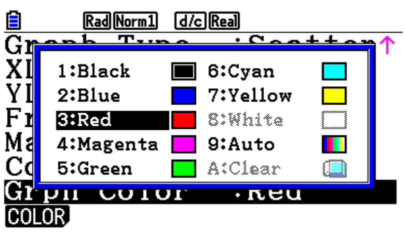


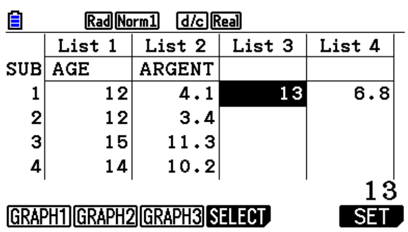


Nous pouvons donc conclure que l'âge moyen du groupe est de 13 ans (et ils s'en écartent, « en moyenne », de 1 an et 7 mois) et s'ils recevaient tous le même montant d'argent de poche, ils aurait chacun 6,8 euros (avec en moyenne une variation de 3 euros).

En encodant les coordonnées du point moyen comme un autre ensemble de données, nous allons pouvoir superposer deux nuages de points : le nuage de points initial, et un nuage de points composé uniquement du point moyen. Définir le nouveau nuage de points dans GRAPH2, en prenant un affichage différent de GRAPH1 (éventuellement en changeant la couleur selon le modèle de la calculatrice).

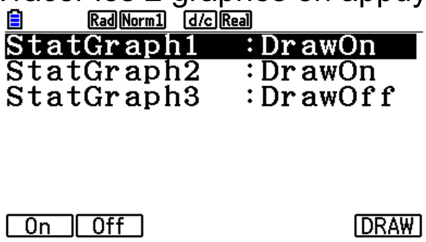


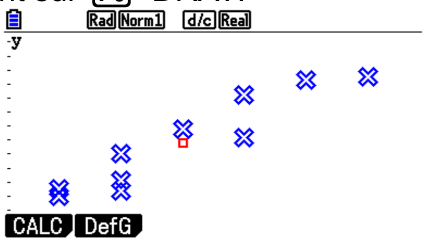




→ Sélectionner F4 SELECT et sélectionner les graphes à afficher simultanément.

→ Tracer les 2 graphes en appuyant sur F6 DRAW





Astuce :
Nous pouvons reprendre les coordonnées du point moyen via les touches
VARS F3 STAT :
pour $\bar{x}$: F1 F2 EXE.
pour $\bar{y}$:
EXIT F2 F1 EXE.

On constate que le point moyen est au « milieu » du nuage.

L'ensemble des couples de données placé dans un repère est appelé ***nuage de points*** ou diagramme de dispersion.

Le nuage de points permet de voir s'il y a ou non une ***orientation*** dûe à un lien entre les abscisses et les ordonnées.

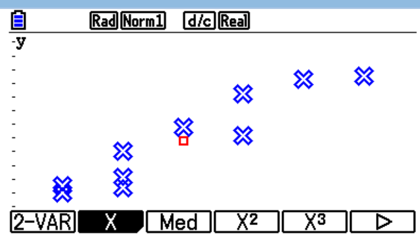
Le ***point moyen*** *ou* ***point milieu*** est le point dont l'abscisse est la moyenne des abscisses et l'ordonnée la moyenne des ordonnées. Il se trouve au "milieu" du nuage.

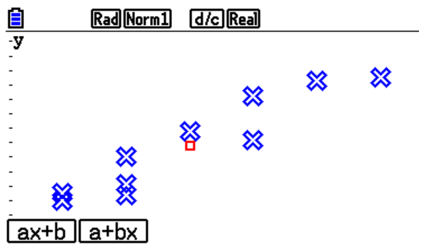
4. CALCULER LES COEFFICIENTS DE LA DROITE DE RÉGRESSION

→ Sélectionner F1 CALC, puis *X* F2 pour le calcul de la droite de régression. Deux options sont proposées en bas de l'écran :

- *ax+b* minimise[2] les écarts verticaux du nuage à la droite.
- *a+bx* minimise les écarts horizontaux du nuage à la droite.

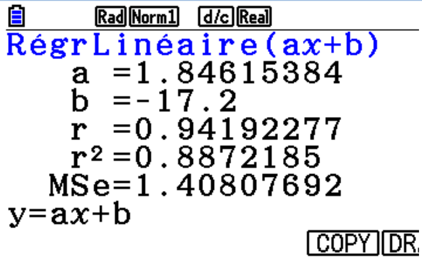
→ Appuyer sur F1





[2] *MSE*, acronyme de Mean Squared Error, ou Erreur Quadratique moyenne, est la moyenne des carrés des écarts entre les ordonnées des points du nuage et les ordonnées sur la droite pour l'abscisse correspondante. ***Attention : ici, il s'agit d'une erreur corrigée.***

- 5 coefficients sont affichés dans l'ordre :
  - La pente de la droite
  - L'ordonnée à l'origine
  - Le coefficient de corrélation
  - Le coefficient de détermination
  - L'erreur quadratique moyenne

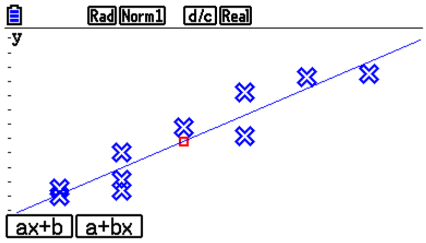


Nous allons bientôt montrer que

- **Ce coefficient de corrélation positif signifie que plus l'adolescent est âgé, plus il a d'argent de poche.**
- **Ce coefficient, comme il est proche de 1, veut dire que le lien entre les variables est très fort.**
- **Le coefficient de détermination traduit la qualité de l'ajustement. Sa valeur, 0,887, signifie que 88,7% des variations d'argent de poche peut être expliqué par l'âge de l'adolescent.**

5. SUPERPOSER NUAGE DE POINTS, POINT MOYEN ET DROITE DE RÉGRESSION

→ Il suffit de sélectionner DRAW [F6]



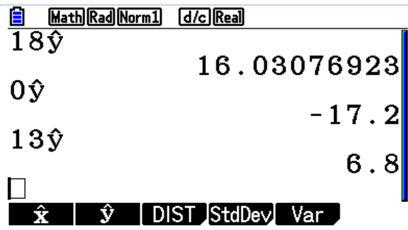
Nous constatons que le point moyen se trouve sur la droite de régression. Non seulement la droite de régression semble bien traduire l'orientation du nuage mais aussi les données s'écartent très peu de la droite.

En conclusion, la droite de régression $y \approx 1.85x - 17.2$ **semble adéquate pour estimer l'argent de poche reçu par un adolescent en fonction de son âge.**

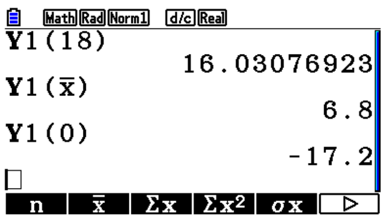
6. PRÉDIRE LES RÉSULTATS

Il existe plusieurs méthodes pour prédire théoriquement un résultat. En voici deux, relativement simples, à utiliser dans le menu RUN.

→ Accéder à l'ordonnée estimée par la droite de régression en tapant la séquence [OPTN] [F5] [F2].



→ … si l'équation de la droite a été précédemment enregistrée via COPY (éventuellement après avoir vérifié la syntaxe de la formule et recalculé les paramètres de dispersion), sélectionner la séquence [VARS] [F4] [F1].



Un jeune de 18 ans recevrait 16,1 euros et un nouveau-né devrait 17,2 euros à ses parents!! L'interprétation du résultat ne peut se faire de manière raisonnable qu'aux alentours du domaine de variation (ici entre 11 et 16 ans). Nous observons au passage que le point moyen appartient bien à la droite puisque $Y1(\overline{x}) = \overline{y}$ **.**

3. Construction intuitive des paramètres de la droite de régression et des coefficients

Comment établir la formule de la droite de régression ? Pourquoi le point moyen appartient-il à la droite de régression ? Que signifie le coefficient de corrélation ? Et celui de détermination ? Pour répondre à ces questions, nous allons avancer pas à pas, à partir de ce que nous connaissons : les statistiques à une variable.

1.1. Changement de repère: point moyen

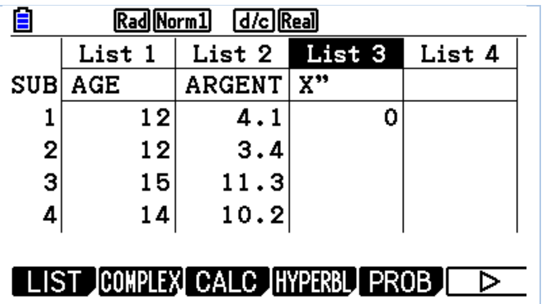
✍ Le point milieu est central au nuage de points. Nous allons donc modifier les axes du graphique de dispersion en plaçant leur origine au centre de gravité du nuage de points. Désignons les coordonnées des points observés dans ce nouveau repère par $x_i'' = x_i - \bar{x}$ et $y_i'' = y_i - \bar{y}$ et déterminons son point moyen.

1	LIST1	LIST2	LIST3	LIST4
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$
1	12	4,1		
2	12	3,4		
3	15	11,3		
4	14	10,2		
5	16	11,5		
6	14	7,2		
7	12	6,0		
8	13	7,8		
9	11	3,5		
10	11	3,0		
Moyenne	**13**	**6,8**		

La calculatrice permet de remplir en une seule opération l'ensemble de la colonne considérée, en encodant la formule correspondante dans la zone des libellés des colonnes (List1, List2, etc.), grâce à la touche OPTN.

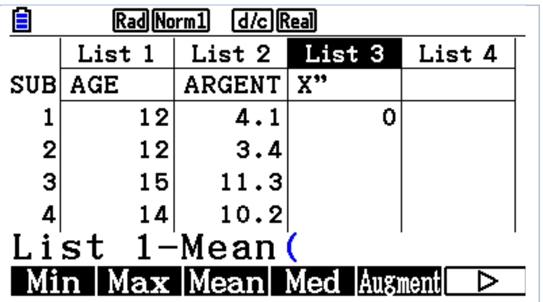
→ Après avoir éventuellement encodé le titre de la colonne dans la cellule SUB, se positionner avec le curseur tout au dessus de la colonne, sur le libellé List3.

→ Taper sur la touche OPTN

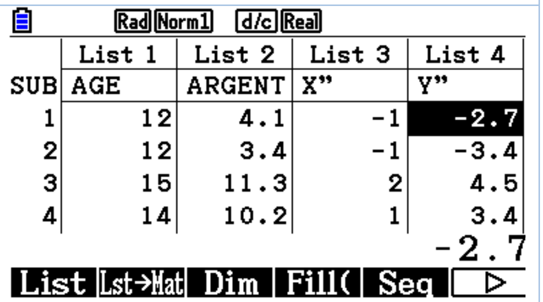


→ Sélectionner [F1] LIST pour sélectionner les opérations disponibles sur les listes

→ Encoder la formule de la troisième colonne $List1 - Mean(List1)$ au moyen de la séquence suivante : [F1] [1] [−] [OPTN] [F1] [F6] [F3] [F6] [F6] [F1] [1] [)] [EXE]



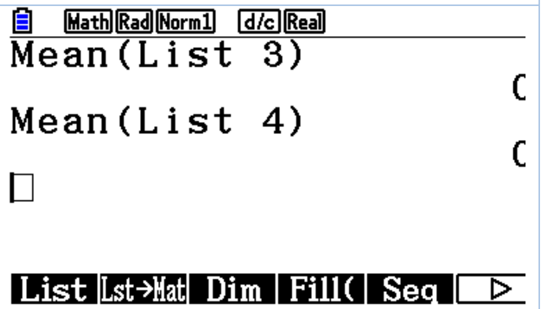
→ Refaire de même pour la colonne suivante ($List2 - Mean(List2)$) au moyen de la séquence suivante : [OPTN] [F1] [F1] [2] [−] [OPTN] [F1] [F6] [F3] [F6] [F6] [F1] [2] [)] [EXE]



→ Pour remplir la dernière ligne du tableau[3], contenant la moyenne de chacune de ces colonnes, aller dans le menu RUN/Exe-Mat

→ [OPTN] [F1] [F6] [F3] [F6] [F6] [F1] [3] [)] [EXE]

→ [F6] [F3] [F6] [F6] [F1] [4] [)] [EXE]



Sans surprise la moyenne est nulle, puisqu'il s'agit de l'écart moyen à la moyenne, qui est toujours nul $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i) - \frac{1}{N}\sum_{i=1}^{N}\bar{x} = \bar{x} - \frac{N\bar{x}}{N} = 0$

Le point moyen de ce nouveau nuage de point est donc l'origine du repère $(0\,;0)$.

Nous pouvons remplir le tableau avec les valeurs centrées en $(\bar{x}\,;\bar{y})$:

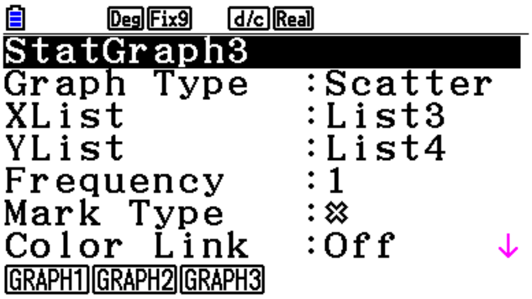
1	LIST1	LIST2	LIST3	LIST4
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$
1	12	4,1	-1	-2,7
2	12	3,4	-1	-3,4
3	15	11,3	2	4,5
4	14	10,2	1	3,4
5	16	11,5	3	4,7
6	14	7,2	1	0,4
7	12	6,0	-1	-0,8
8	13	7,8	0	1
9	11	3,5	-2	-3,3
10	11	3,0	-2	-3,8
Total/10	**13**	**6,8**	**0**	**0**

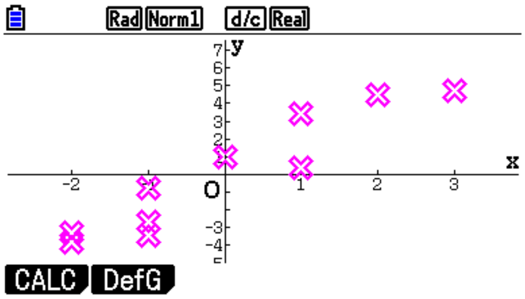
[3] Conseil : à la calculatrice, ne pas calculer la moyenne en dessous de la colonne, car le résultat sera considéré ensuite comme faisant partie des données.

1.2. ORIENTATION DU NUAGE : COVARIANCE

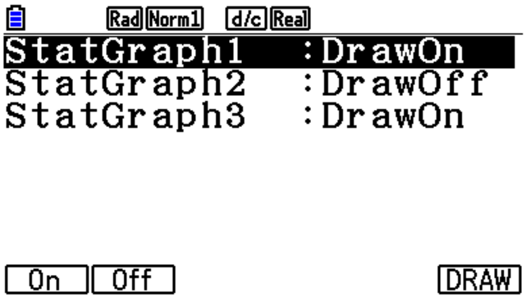
✍ L'orientation du nouveau nuage de points, translaté au point milieu, est-elle modifiée ?

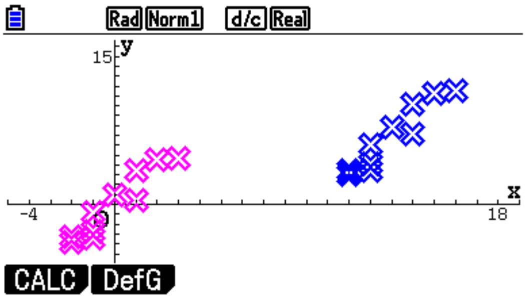
En assignant le nouveau nuage de points à GRAPH3, on constate que le nuage de points déplacé est bien centré en $(0;0)$.





En superposant les deux nuages de points pour les comparer, nous constatons que le nuage translaté conserve la même orientation.



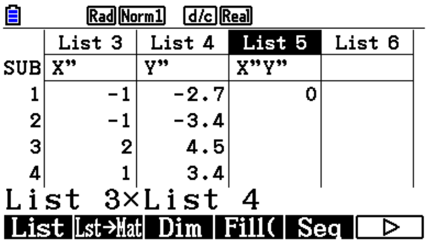


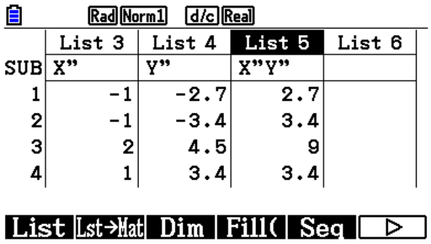
La translation n'a pas modifié l'orientation.

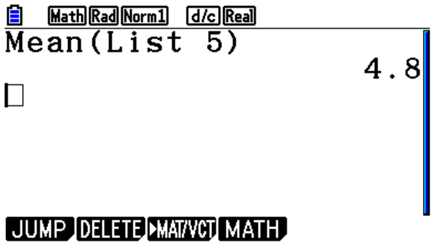
Dans l'univers papier-crayon, il n'est pas utile de translater les données, il suffit de déplacer le repère en le centrant au point moyen. Il est alors évident que l'orientation du nuage n'est pas modifiée par la translation.

✍ En définissant les quatre quadrants définis par les axes du nouveau repère, dans quels quadrants se trouvent les nouvelles observations ? Et comment résumer ceci en une information ?

Les observations se trouvent majoritairement dans les premier et troisième quadrants, autrement dit les abscisses et ordonnées de chaque point **sont de même signe.** Vérifions cela algébriquement en complétant le tableau :







1	LIST1	LIST2	LIST3	LIST4	LIST5
i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
1	12	4,1	-1	-2,7	2,7
2	12	3,4	-1	-3,4	3,4
3	15	11,3	2	4,5	9,0
4	14	10,2	1	3,4	3,4
5	16	11,5	3	4,7	14,1
6	14	7,2	1	0,4	0,4
7	12	6,0	-1	-0,8	0,8
8	13	7,8	0	1	0
9	11	3,5	-2	-3,3	6,6
10	11	3,0	-2	-3,8	7,6
Moyenne	**13**	**6,8**	**0**	**0**	**4,8**

La moyenne de ces produits est positive car, même si quelques points se trouvent dans les autres quadrants, les abscisses et ordonnées de la plupart des points du nuage sont de même signe.

La moyenne de ces produits est appelée ***covariance*** et notée s_{xy}.

Elle possède les mêmes propriétés que la variance à l'exception du fait qu'elle peut être positive ou négative.

Une covariance **positive** correspond à des variations des deux variables **dans le même sens**. Inversement une covariance négative correspond à des variations en sens opposé.

1.3. PONDÉRATION ET ACCROISSEMENT MOYEN : DROITE ORIENTÉE DANS LE NUAGE

Nous allons construire l'équation d'une droite qui traduirait l'orientation générale du nuage de points, de manière intuitive, nous l'appellerons droite de régression. Nous verrons ensuite qu'elle est proportionnelle à la covariance s_{xy} (qui, nous l'avons déjà remarqué, traduit bien l'orientation du nuage). Nous montrerons plus tard, algébriquement, que c'est la « meilleure ».

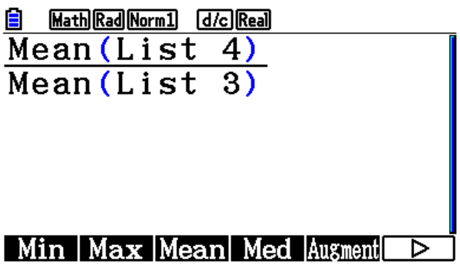
✍ La pente de la droite de régression devrait représenter l'accroissement moyen d'argent de poche, $a = \dfrac{\Delta y_{moyen}}{\Delta x_{moyen}}$. Comment définir ces $\Delta y_{moyen}, \Delta x_{moyen}$?

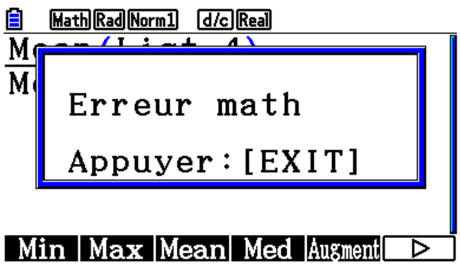
A ce stade, il est intéressant de laisser libre cours à l'imagination des élèves et les laisser proposer des solutions originales.

Souvent, ils proposent spontanément de prendre les plus grands écarts en abscisses et en ordonnées. On peut leur montrer que non seulement il faut trier tous les points et les afficher pour déterminer les points extrêmes mais aussi, et surtout, cela donne beaucoup de poids aux valeurs extrêmes, alors que celles-ci peuvent complètement dévier la droite. En général, ils proposent alors de calculer une moyenne, moyenne sur les ordonnées, moyenne sur les abscisses.

✍ Une telle pente $a = \dfrac{\text{déviation moyenne des ordonnées}}{\text{déviation moyenne des abscisses}} = \dfrac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)/n}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)/n}$ pourrait-elle convenir ?

Comme nous l'avons déjà vu plus haut, l'écart moyen est toujours nul, cette expression est impossible à calculer car nous obtenons 0/0.





✍ Comment modifier les numérateur et dénominateur pour qu'ils ne s'annulent pas ?

Certains élèves proposent de travailler en valeur absolue, ou d'élever au carré, mais l'orientation d'une telle droite serait toujours positive, ce qui n'est pas ce que nous voulons.

Par contre, nous pouvons leur faire remarquer qu'en pondérant les observations de manière appropriée, nous pouvons en même temps éviter cette division par 0 et traduire l'orientation du nuage.

$$a=\frac{\sum_{i=1}^{n} w_i\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n} w_i\left(x_i-\bar{x}\right)}$$ où les w_i sont les poids de chaque observation.

✍ Que prendre alors comme poids ? Dans le tableau de résultats, quels sont les points qui influent le moins sur l'orientation du nuage? Et ceux qui sont prépondérants ?

On constate que les points proches du point moyen apportent peu d'informations alors que les points éloignés modifient sensiblement l'orientation du nuage. Prendre des poids $w_i=x_i-\bar{x}$ ou $w_i=y_i-\bar{y}$ revient à décider si c'est l'abscisse qui détermine l'ordonnée ou l'inverse. Comme nous essayons d'estimer les ordonnées à partir des abscisses, nous donnons plus d'importance aux points dont les abscisses sont éloignées de la moyenne.

✍ Que devient alors la pente avec de tels poids ?

$$a=\frac{\sum_{i=1}^{n} w_i\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n} w_i\left(x_i-\bar{x}\right)}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$$

✍ Peut-on transformer cette expression en fonction de la covariance, qui, nous l'avons déjà remarqué, traduit l'orientation du nuage ?

Rappelons que la covariance $s_{xy}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$ que l'on reconnaît au numérateur à un facteur n près. En divisant par le nombre total d'observations au numérateur et au dénominateur, nous ne modifions pas l'expression. Nous obtenons donc

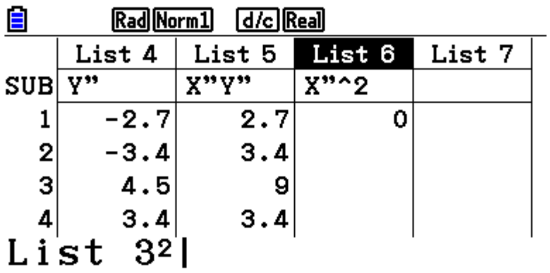
$$a=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)/n}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2/n}$$

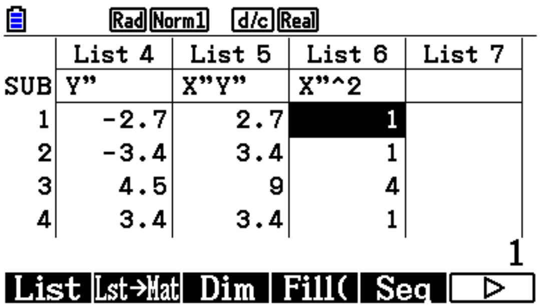
Nous retrouvons un indicateur de dispersion bien connu au dénominateur : la variance des abscisses $s_x{}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$.

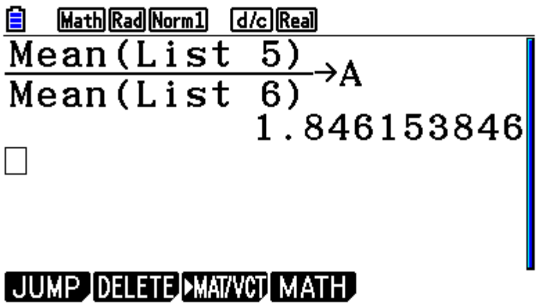
Nous pouvons donc écrire plus simplement cette pente : $a=\frac{s_{xy}}{s_x{}^2}$

✍ Que vaut la pente de la droite de notre nuage de points?

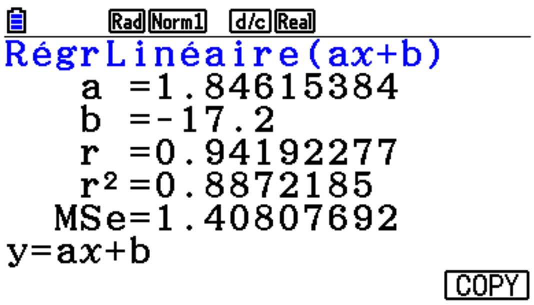
Après avoir encodé le calcul des carrés des écarts des abscisses dans la dernière colonne, nous pouvons déterminer la valeur de la pente dans le menu RUN puisqu'il suffit de prendre la moyenne de cette dernière pour obtenir la variance.





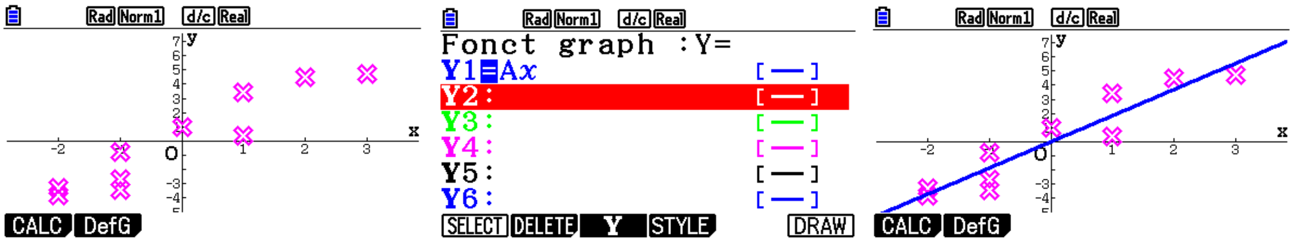


C'est bien ce que la calculatrice avait affiché lors du calcul des paramètres de régression :



✍ Vérifions avec la calculatrice qu'une telle pente traduit bien l'orientation du nuage.

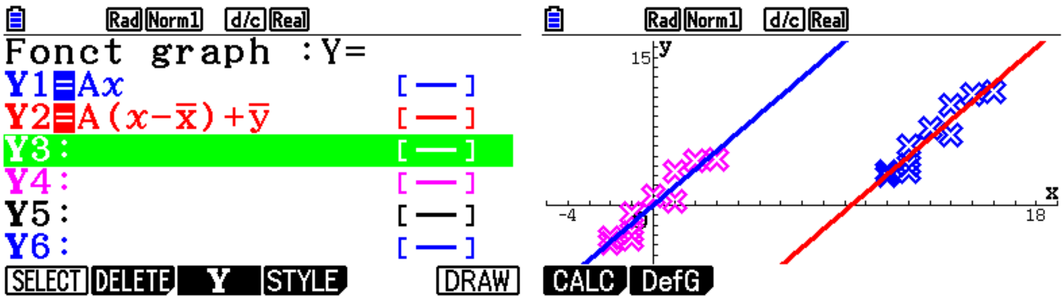
Considérons la droite passant par l'origine et de pente a et superposons-la au nuage de points (GRAPH3) via la touche DefG F2 :



Cette droite traduit bien l'orientation du nuage, autrement dit, elle permet de prédire l'argent de poche pour un âge donné.

✍ Déterminer l'équation de la droite translatée dans le repère initial et la superposer au nuage de points initial.

Translatons cette droite dans l'autre repère, par manipulation d'équation en utilisant les coordonnées du point moyen et affichons les deux nuages de points :



L'équation de la droite $y = a(x - \overline{x}) + \overline{y}$. **Par construction**, elle passe par le point moyen.

Une droite contenant le point moyen passe dans le nuage de points.

Une droite dont la pente vaut $a = s_{xy}/s_x^2$ a la même orientation que le nuage de points.

1.4. CHANGEMENT D'ÉCHELLE: COEFFICIENT DE CORRÉLATION

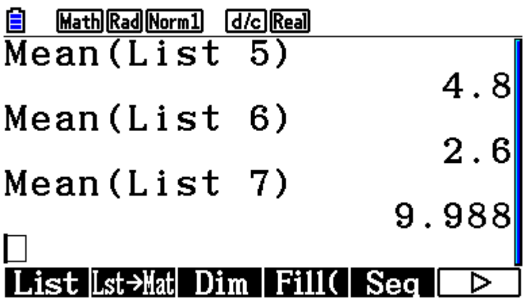
La covariance traduit l'orientation du nuage. Cependant son interprétation est gênée par la dépendance vis-à-vis des unités choisies. En effet, un même nuage de points exprimé dans telle ou telle unité, aura la même orientation mais aura une covariance différente. Nous allons donc "standardiser" (ou "normaliser" ou "réduire") cette covariance, en divisant les observations correspondantes par leur écart-type (racine carrée de la variance, de même unité que les observations et traduisant l'écart moyen).

✍ Désignons les coordonnées des points observés dans ce nouveau repère par $u_i = \dfrac{x_i - \bar{x}}{s_X}; v_i = \dfrac{y_i - \bar{y}}{s_Y} \quad (i = 1..n)$ et complétons le tableau pour voir comment se comporte la covariance « normalisée ».

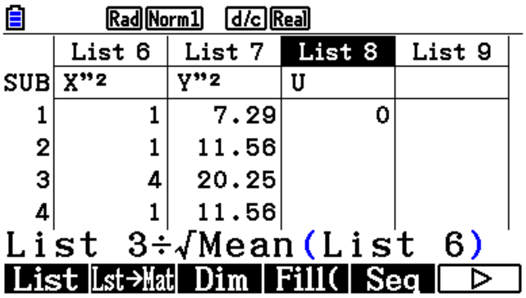
Pour obtenir l'écart-type des ordonnées, nous avons besoin de calculer, comme nous l'avons déjà fait pour les abscisses, la variance de celles-ci, moyenne des carrés des écarts à leur moyenne respective.

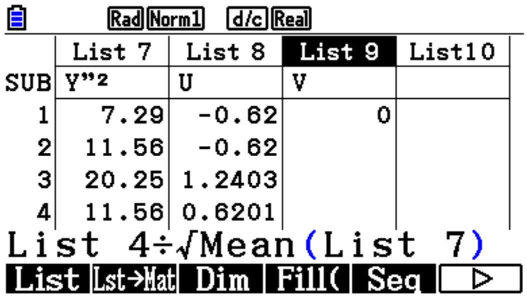
1	LIST1	LIST2	LIST3	LIST4	LIST5	LIST6	LIST7	LIST8	LIST9	LIST10
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}).(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\frac{x_i - \bar{x}}{s_x} . \frac{y_i - \bar{y}}{s_y}$
1	12	4,1	-1	-2,7	2,7	1	7 ,29			
2	12	3,4	-1	-3,4	3,4	1	11,56			
3	15	11,3	2	4,5	9,0	4	20,25			
4	14	10,2	1	3,4	3,4	1	11,56			
5	16	11,5	3	4,7	14,1	9	22,09			
6	14	7,2	1	0,4	0,4	1	0,16			
7	12	6,0	-1	-0,8	0,8	1	0,64			
8	13	7,8	0	1	0	0	1,00			
9	11	3,5	-2	-3,3	6,6	4	10,89			
10	11	3,0	-2	-3,8	7,6	4	14,44			
Moyenne	**13**	**6,8**	**0**	**0**	**4,8**					

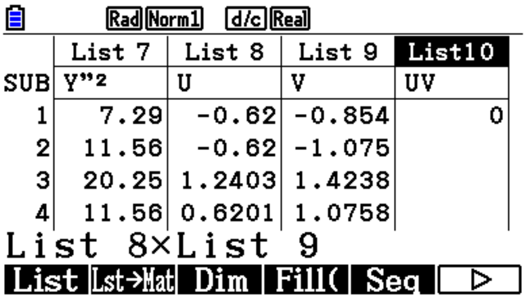
Il suffit de calculer les moyennes dans le menu RUN pour obtenir la variance.

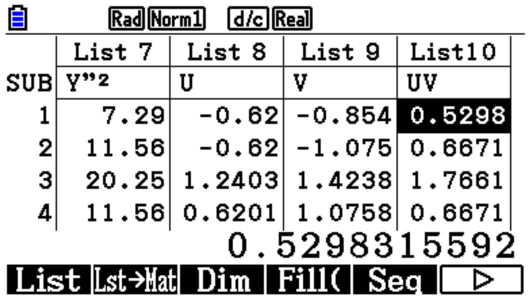


Nous prenons la racine des variances, moyennes des listes 6 et 7, pour obtenir s_x et s_y et compléter le tableau des coordonnées « normalisées ».



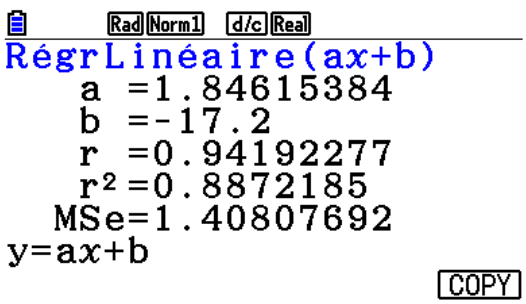






1	LIST1	LIST2	LIST3	LIST4	LIST5	LIST6	LIST7	LIST8	LIST9	LIST10
i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\frac{x_i-\bar{x}}{s_X}$	$\frac{y_i-\bar{y}}{s_Y}$	$\frac{(x_i-\bar{x})\cdot(y_i-\bar{y})}{s_X\cdot s_Y}$
1	12	4,1	-1	-2,7	2,7	1	7 ,29	-0,62	-0,85	0,53
2	12	3,4	-1	-3,4	3,4	1	11,56	-0,62	-1,07	0,67
3	15	11,3	2	4,5	9,0	4	20,25	1,24	1,42	1,77
4	14	10,2	1	3,4	3,4	1	11,56	0,62	1,08	0,67
5	16	11,5	3	4,7	14,1	9	22,09	1,86	1,48	2,77
6	14	7,2	1	0,4	0,4	1	0,16	0,62	0,13	0,08
7	12	6,0	-1	-0,8	0,8	1	0,64	-0,62	-0,25	0,16
8	13	7,8	0	1	0	0	1,00	0	0,32	0
9	11	3,5	-2	-3,3	6,6	4	10,89	-1,24	-1.04	1,30
10	11	3,0	-2	-3,8	7,6	4	14,44	-1,24	-1,20	1,49
Moyenne	**13**	**6,8**	**0**	**0**	**4,8**	**2,6**	**9,99**	**0**	**0**	**0,94**

La moyenne de ces produits, covariance centrée et réduite, est appelée ***coefficient de corrélation*** et notée r. C'est bien ce que la calculatrice a déterminé :



$r = \dfrac{s_{xy}}{s_x s_y}$. Le coefficient de corrélation est donc une covariance sans unité.

On peut également lier la pente de la droite et ce coefficient.

En effet, puisque

$$
\begin{aligned}
r &= \frac{s_{xy}}{s_x s_y} \\
&= \left(\frac{s_{xy}}{s_x s_y}\right) \cdot \frac{s_x}{s_x} \\
&= \frac{s_{xy}}{s_x{}^2} \cdot \frac{s_x}{s_y} \\
&= a \frac{s_x}{s_y}
\end{aligned}
$$

Nous pouvons donc conclure que $a = r\dfrac{s_y}{s_x}$.

La pente de la droite est en quelque sorte le coefficient de corrélation remis à échelle.

Le coefficient de corrélation traduit le lien entre les abscisses et les ordonnées, sans unité.

Le coefficient de corrélation, noté ***r***, vaut la covariance standardisée: $s_{xy} / (s_x . s_y)$

La pente de la droite de régression $a = r . s_y / s_x$ est le coefficient de corrélation "remis à échelle".

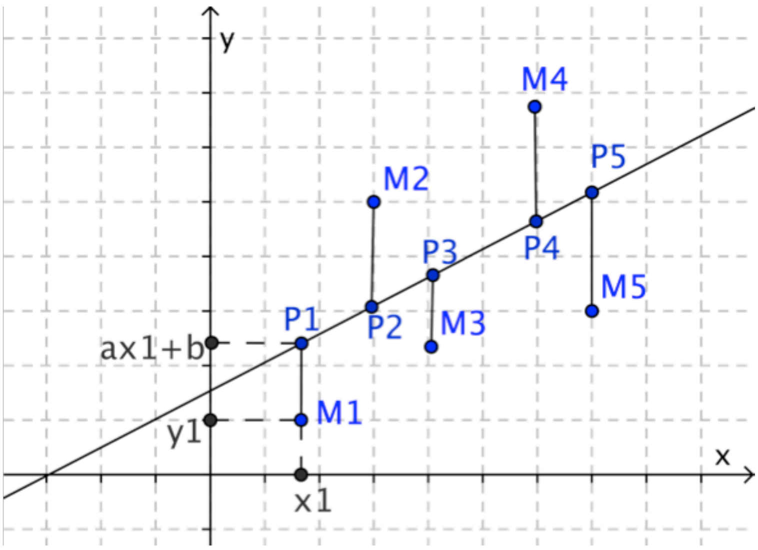
1.5. MEILLEURE DROITE D'AJUSTEMENT: ERREUR QUADRATIQUE MOYENNE

Au vu du graphique de dispersion, le montant d'argent de poche attribué aux dix jeunes semble être fonction de leur âge. Il est raisonnable de penser que cette dépendance des ordonnées par rapport aux abscisses peut être décrite par un modèle du type $y = ax + b$ [4], les ordonnées théoriques étant fonction des abscisses.

Un bon indicateur de la qualité d'ajustement serait un indicateur qui permet de déterminer à quel point l'équation de la droite est adaptée pour décrire la distribution des points, autrement dit à quel point les points sont éloignés ou pas de la droite calculée.

Pour cela, définissons les ***résidus***, écarts entre l'ordonnée observée et l'ordonnée théorique: $e_i = y_i - (ax_i + b)$.

Sur le graphe ci-contre, à chaque point du nuage de points, on peut faire correspondre un point de la droite "idéale", correspondant à la loi sous-jacente ayant la même abscisse que le point considéré. Les écarts entre les points du nuage de points et les points « idéaux » représentent donc les résidus.

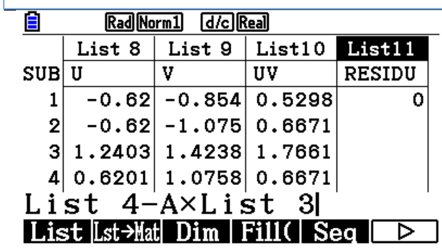


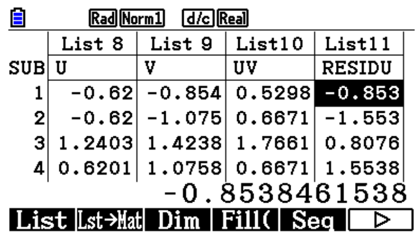
Résidu e_1=écart entre $M_1(x_1\,;y_1)$ et $P_1(x_1\,;ax_1+b)$

Plus les points sont proches de la droite, plus les écarts seront faibles.

Intuitivement, on voudrait imposer que la somme des résidus soit la plus petite possible.

✍ Regardons les écarts par rapport à la droite et déterminons sa moyenne.





[4] On suppose donc que les variables *X* et *Y* ne jouent pas un rôle symétrique: la variable *Y* est dite *dépendante* (ou expliquée) et *X* est appelée variable *explicative*.

Ce critère n'est pas assez fort car la moyenne des résidus de toute droite passant par le point moyen, indépendamment de sa pente, est **toujours nulle**.

En effet

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} e_i &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)\\ &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb\right)\\ &= \overline{y}-a\overline{x}-b\\ &= 0\end{aligned}$$

Nous allons donc imposer un critère encore plus fort : plus la moyenne des carrés des résidus est faible, plus la droite est ajustée au nuage de points. La droite la mieux ajustée au nuage de points est donc celle dont la moyenne des carrés des résidus est la plus petite.

Pour déterminer a,b, nous allons donc minimiser $\frac{1}{n}\sum_{i=1}^{n} e_i^{\,2}$. D'où le nom de ***méthode des moindres carrés***.

La droite d'équation $y = ax+b$ qui minimise cette moyenne des carrés des résidus est appelée ***droite de régression*** et cette moyenne des carrés des résidu est appelée ***erreur quadratique moyenne***, notée *MSE* (acronyme de Mean Squared Error).

Le ***résidu*** d'une observation est l'écart entre l'ordonnée observée et l'ordonnée théorique.

La moyenne des résidus d'une droite passant par le point moyen est toujours nulle.

La moyenne des carrés des résidus minimale est appelée ***erreur quadratique moyenne***, notée aussi MSE.

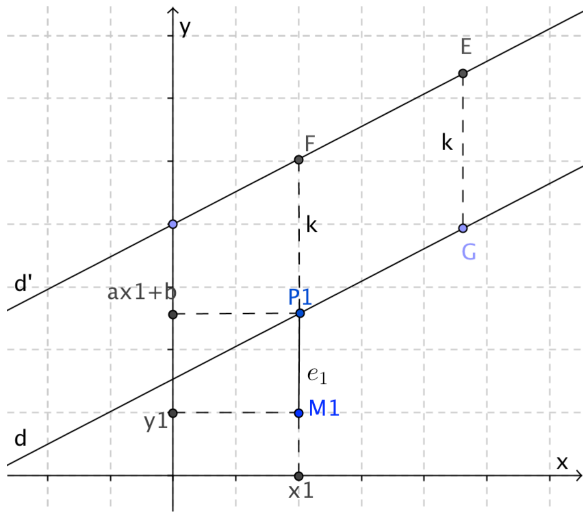
La droite correspondant à l'erreur quadratique moyenne est celle qui s'ajuste "le mieux" au nuage de point. Cette droite est appelée ***droite de régression.***

Intuitivement, nous avions pressenti que la droite qui s'ajuste le mieux aux données doit passer par le point milieu et être proportionnelle à la covariance.

Montrons algébriquement, en deux étapes distinctes, qu'il ne peut en être autrement, à partir de l'erreur quadratique moyenne.

Tout d'abord, pour une pente donnée, la droite qui minimise la moyenne quadratique des résidus (autrement dit la droite de régression) passe nécessairement par le point milieu.

En effet, n'importe quelle autre droite d', de même pente et translatée de k unités verticalement, aura sa moyenne quadratique des résidus e_i' supérieure à celle des résidus de la droite passant par le point milieu, les résidus étant eux aussi augmentés de k unités.



Droite de même pente translatée de k unités verticalement:

Le résidu de d' vaut le résidu de d augmenté de k unités.

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} e_i'^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(e_i+k\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{1}{n}\sum_{i=1}^{n} k^2 + \frac{1}{n}\sum_{i=1}^{n} 2ke_i \\
&= MSE + k^2 + 2k\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right) \\
&= MSE + k^2 + 0 \\
&\geq MSE + k^2
\end{aligned}
$$

Ensuite, minimiser la somme des carrés des écarts revient à minimiser une fonction polynomiale de degré 2 de variable a.

En effet, puisque l'équation de la droite de régression passe par le point milieu, elle s'écrit donc $y = a(x-\overline{x})+\overline{y}$ et nous pouvons exprimer les résidus en fonction de cette équation.

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(a(x_i-\overline{x})+\overline{y}\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i-\overline{y}\right) - a\left(x_i-\overline{x}\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i-\overline{y}\right)^2 - 2a\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right) + a^2\left(x_i-\overline{x}\right)^2\right) \\
&= s_y{}^2 - 2as_{xy} + a^2 s_x{}^2
\end{aligned}
$$

Il s'agit bien d'une fonction polynomiale de degré 2 de variable a.

Le minimum MSE est de cette fonction est atteint quand la variable $a = \frac{s_{xy}}{s_x^{\,2}}$ et vaut alors

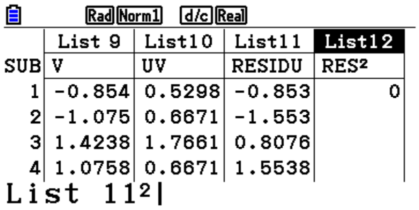
$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n} e_i^{\,2} \\ &= s_y^{\,2} - 2\frac{s_{xy}}{s_x^{\,2}} s_{xy} + \left(\frac{s_{xy}}{s_x^{\,2}}\right)^2 s_x^{\,2} \\ &= s_y^{\,2} - \left(\frac{s_{xy}}{s_x}\right)^2 \end{aligned}.$$

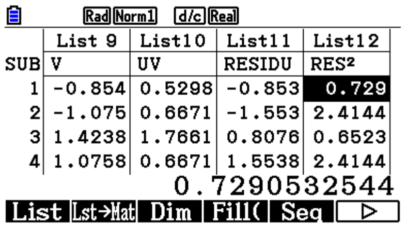
La droite de régression passe **toujours** par le point milieu.

La pente de la droite de régression $a = s_{xy} / s_x^{\,2}$

L'erreur quadratique moyenne $MSE = s_y^{\,2} - (s_{xy}/s_x)^2$

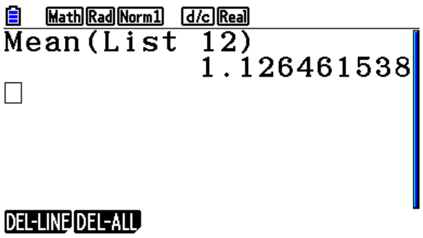
✍ Calculons la moyenne des carrés des écarts de notre droite et comparons-la avec celle affichée par la calculatrice.

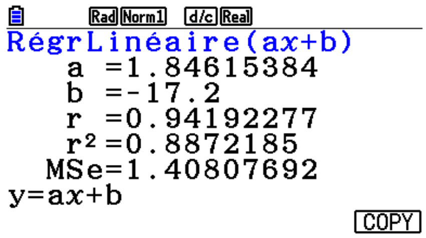




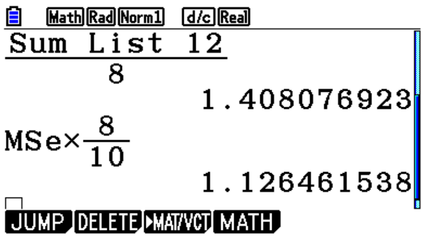
1	LIST1	LIST2	LIST3	LIST4	…	LIST11	LIST12
i	x_i	y_i	$x_i-\overline{x}$	$y_i-\overline{y}$	…	e_i	e_i^2
1	12	4,1	-1	-2,7		-0,85	0.73
2	12	3,4	-1	-3,4		-1,55	2,41
3	15	11,3	2	4,3		0,80	0,65
4	14	10,2	1	3,4		1.55	2,42
5	16	11,5	3	4,7		-0,84	0,70
6	14	7,2	1	0,4		-1,45	2,09
7	12	6,0	-1	-0,8		1,05	1,09
8	13	7,8	0	1		1	1
9	11	3,5	-2	-3,3		0,39	0,15
10	11	3,0	-2	-3,8		-0,11	0,01
Moyenne	**13**	**6,8**	**0**	**0**		**0**	**1,13**

Notre erreur quadratique calculée est différente de celle affichée par la calculatrice.





La calculatrice affiche en fait une erreur quadratique corrigée (car 2 degrés de liberté : les paramètres a et b à estimer) $MSE \text{ corrigée} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)$. On peut donc retrouver l'erreur quadratique moyenne non corrigée en multipliant par $\frac{n-2}{n}$.

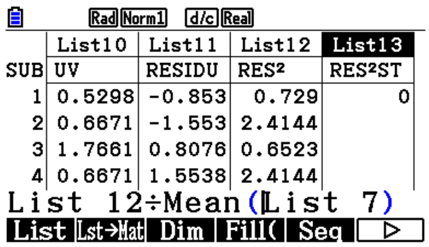


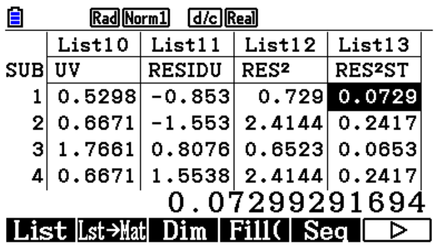
1.6. QUALITÉ DE L'AJUSTEMENT : COEFFICIENT DE DÉTERMINATION

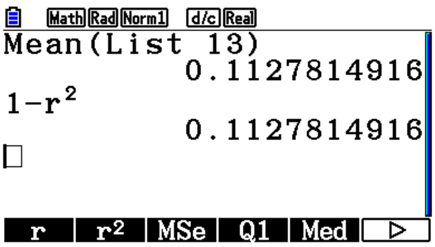
L'erreur quadratique moyenne va de pair à la droite de régression qui est la droite qui s'ajuste le mieux au nuage de points. Tout comme nous avons standardisé la covariance, nous allons standardiser l'erreur quadratique moyenne.

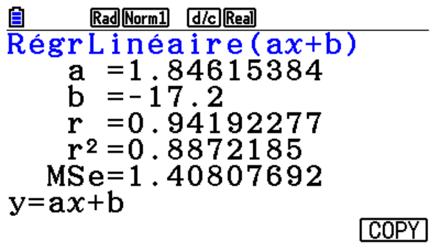
✍ Calculons cette erreur quadratique moyenne standardisée en posant $e_i'^2 = \frac{e_i^2}{s_x^2}$ et comparons-la avec la valeur r^2 affichée par la calculatrice.

1	LIST1	LIST2	LIST3	LIST4	…	LIST11	LIST12	LIST13
i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	…	e_i	e_i^2	e_i^2 / s_Y^2
1	12	4,1	-1	-2,7		-0,85	0.73	0.07
2	12	3,4	-1	-3,4		-1,55	2,41	0,24
3	15	11,3	2	4,3		0,80	0,65	0,06
4	14	10,2	1	3,4		1.55	2,42	0,24
5	16	11,5	3	4,7		-0,84	0,70	0,07
6	14	7,2	1	0,4		-1,45	2,09	0,21
7	12	6,0	-1	-0,8		1,05	1,09	0,11
8	13	7,8	0	1		1	1	0,1
9	11	3,5	-2	-3,3		0,39	0,15	0,01
10	11	3,0	-2	-3,8		-0,11	0,01	0,00
Moyenne	**13**	**6,8**	**0**	**0**		**0**	**1,13**	**0,11**









L'erreur quadratique moyenne standardisée vaut $1 - r^2$.

Nous pouvons d'ailleurs le montrer algébriquement :

$$\begin{aligned}\frac{MSE}{s_y^{\ 2}} &= \frac{\frac{1}{n}\sum_{i=1}^{n} e_i^{\ 2}}{s_y^{\ 2}} \\ &= \frac{s_y^{\ 2} - \left(\frac{s_{xy}}{s_x}\right)^2}{s_y^{\ 2}} \\ &= 1 - \left(\frac{s_{xy}}{s_x}\right)^2 \\ &= 1 - r^2\end{aligned}$$

✍ Comment interpréter ce lien entre erreur quadratique moyenne standardisée et coefficient de corrélation?

En se rappelant que la variance $s_x^{\ 2}$ d'un échantillon est sensible aux changements d'échelle et insensible aux translations[5], nous pouvons réécrire $a^2 s_x^{\ 2} = s_{ax+b}^{\ \ 2}$. Ceci va nous permettre d'exprimer $s_y^{\ 2}$, la variabilité des ordonnées, autrement puisque $MSE = s_y^{\ 2} - \left(\frac{s_{xy}}{s_x}\right)^2$, qui peut aussi s'écrire en fonction de a $MSE = s_y^{\ 2} - \left(a s_x\right)^2$

$$s_y^{\ 2} = \left(a s_x\right)^2 + MSE$$

$$s_y^{\ 2} = a^2 s_x^{\ 2} + MSE$$

$$\underbrace{s_y^{\ 2}}_{\substack{\text{variabilité}\\ \text{de l'ordonnée}}} = \underbrace{s_{ax+b}^{\ \ 2}}_{\substack{\text{variabilité}\\ \text{due à l'abscisse}}} + \underbrace{MSE}_{\substack{\text{variabilité}\\ \text{non expliquée}}} \qquad (1)$$

Autrement dit, la variabilité de l'ordonnée est obtenue par deux quantités complémentaires : une variabilité due à la dépendance linéaire de l'ordonnée par rapport à l'abscisse et une variabilité résiduelle, due à des fluctuations non expliquées par la droite.

5

$$s_{ax} = \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\overline{x})^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2 s_x$$

$$s_{x+b} = \frac{1}{n}\sum_{i=1}^{n}((x_i + b) - (\overline{x} + b))^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^{\ 2}$$

En définissant la ***qualité de l'ajustement*** comme le rapport entre la variabilité due à l'abscisse et la variabilité de l'ordonnée, nous partir de l'expression (1) , en divisant par s_y^2 , que

$$\frac{s_y^2}{s_y^2} = \frac{s_{ax+b}^2}{s_y^2} + \frac{MSE}{s_y^2}$$

$$1 = \text{Qualité d'ajustement} + \frac{MSE}{s_y^2}$$

$$\begin{aligned}\text{Qualité d'ajustement} &= 1 - \frac{MSE}{s_y^2} \\ &= 1-(1-r^2) \\ &= r^2\end{aligned}$$

Cette qualité d'ajustement, r^2 , appelée **coefficient de détermination,** représente donc le pourcentage de la variance s_y^2 imputable à la dépendance linéaire de y en x et varie entre 0% et 100%.

Si la droite est parfaitement ajustée au nuage de points, l'erreur quadratique moyenne est nulle et donc la qualité d'ajustement vaut 100%.

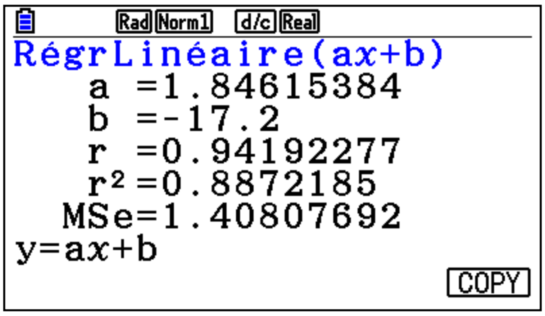
Plus la qualité de l'ajustement est mauvaise, plus l'erreur quadratique est élevée, plus sa valeur est proche de 0.

Le ***coefficent de détermination***, r^2, compris entre 0 et 1, représente la qualité d'ajustement de la droite au nuage de points.

Plus r^2 est proche de 1, plus la droite s'ajuste au nuage.

$r^2 = 1 - MSE/ s_y^2$ peut être interprété comme le pourcentage de variablilité de l'ordonnée due à la dépendance linéaire de l'ordonnée par rapport à la valeur de l'abscisse.

✍ Que vaut la qualité de l'ajustement de notre droite de régression aux données ? Et comment l'interpréter ?



Le coefficient de détermination vaut 0.889. La qualité d'ajustement vaut donc 0,887, ce qui signifie que 88,7% des variations d'argent de poche peuvent être directement expliquées par l'âge de l'adolescent.

4. Utilité de la vérification des hypothèses

S'il est tentant de juger de la qualité d'un ajustement au moyen du coefficient de corrélation, il faut rester conscient que des corrélations identiques peuvent provenir de données totalement différentes.

Certaines données aberrantes peuvent fausser complètement le résultat du calcul des coefficients de régression ou le modèle peut tout simplement ne pas être linéaire.

Si les écarts par rapport aux hypothèses du modèle sont faibles, les résultats ne sont pas erronés. Par contre, les conclusions peuvent n'avoir aucun sens si le modèle n'est pas vérifié.

Pour illustrer ces différents cas, voici 5 ensemble de 16 couples de valeurs[6] dont les nuages de points respectifs sont très différents alors que leurs moyennes respectives, les coefficients de leur droite de régression et leurs coefficients de détermination sont à peu de choses près IDENTIQUES !

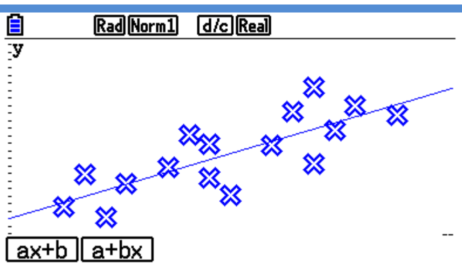
Obs.	X(A) X(B) X(C) X(D)	Y(A)	Y(B)	Y(C)	Y(D)	X(E)	Y(E)
1	7,000	5,535	0,113	7,399	3,864	13,715	5,654
2	8,000	9,942	3,770	8,546	4,942	13,715	7,072
3	9,000	4,249	7,426	8,468	7,504	13,715	8,491
4	10,000	8,656	8,792	9,616	8,581	13,715	9,909
5	12,000	10,737	12,688	10,685	12,221	13,715	9,909
6	13,000	15,144	12,889	10,607	8,842	13,715	9,909
7	14,000	13,939	14,253	10,529	9,919	13,715	11,327
8	14,000	9,450	16,545	11,754	15,860	13,715	11,327
9	15,000	7,124	15,620	11,676	13,967	13,715	12,746
10	17,000	13,693	17,206	12,745	19,092	13,715	12,746
11	18,000	18,100	16,281	13,893	17,198	13,715	12,746
12	19,000	11,285	17,647	12,590	12,334	13,715	14,164
13	19,000	21,365	14,211	15,040	19,761	13,715	15,582
14	20,000	15,692	15,577	13,737	16,382	13,715	15,582
15	21,000	18,977	14,652	14,884	18,945	13,715	17,001
16	**23,000**	**17,690**	**13,497**	**29,431**	**12,187**	**33,281**	**27,435**

Ces cinq ensembles de couples de données, téléchargeables sur le site de Casio éducation dans la rubrique Profs, donnent les mêmes résultats numériques alors qu'ils correspondent à des réalités très différentes quand nous affichons le nuage de points. Pour chacun de ces groupes de points, les paramètres sont à peu près identiques : point moyen valant à peu de chose près (14.9 ; 12.6) avec des écarts types respectifs de 4.7 et 5.03, avec une droite de régression proche de y= 0.8x+0.5 et un coefficient de corrélation proche

[6] Couple de valeurs basés sur le « quatuor d'Anscombe ». ANSCOMBE (1918-2001), mathématicien et statisticien, est connu pour ses études des propriétés des résidus de la régression linéaire. Il crée en 1973 le "Anscombe quartet" (ou "Quatuor d'Anscombe") constitué de quatre ensembles de 11 couples de données, qui ont des statistiques identiques (moyenne, variance, droite de régression et coefficient de corrélation) alors que leurs nuages de points respectifs sont très différents. Il démontre ainsi à la fois l'importance de visualiser les données avant de les analyser et l'effet des données aberrantes (ou outliers) sur les propriétés statistiques. Depuis de nombreux autres ensembles de données ont été créés et sont disponibles dans la littérature ou sur le net. Ce jeu-ci est extrait du manuel de TOMASSONE, intitulé *La régression. Nouveaux regards sur une ancienne méthode statistique* & AL. INRA et Masson, Paris 1992.

de 77%, autrement dit, pas trop mauvais. Et pourtant voici les nuages de points correspondants !!!!!

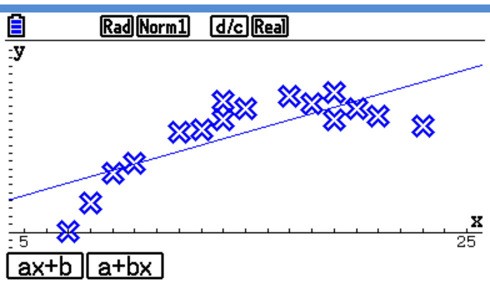
X(A)-Y(A)



La droite calculée semble traduire raisonnablement l'orientation du nuage.

Le modèle paraît valide.

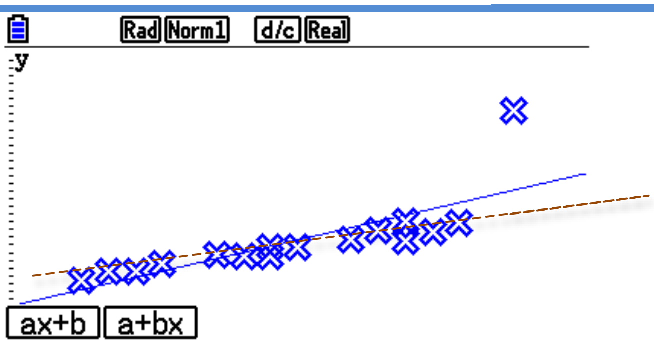
X(B)-Y(B)



La fonction qui s'ajuste au mieux à ce nuage de point semble quadratique.

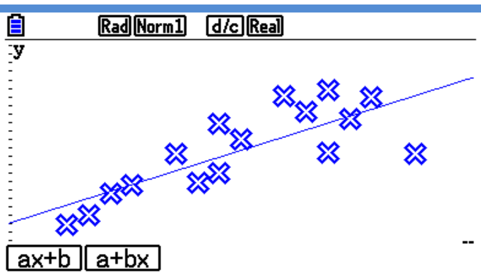
Le modèle est incorrect.

X(C)-Y(C)



La droite ne s'adapte pas bien aux données : beaucoup de points sont situés en dessous de la droite. Ces observations semblent alignées autour d'une autre droite (ajoutée ici en pointillé) dont les paramètres seraient différents. Une seule observation semble en être la cause. **Donnée suspecte à vérifier donc.**

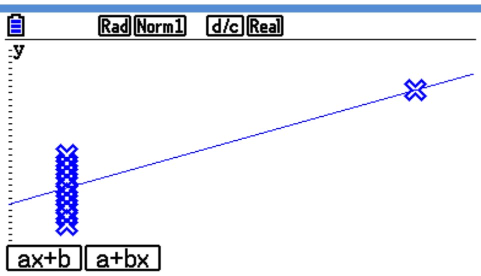
X(D)-Y(D)



Les points se resserrent autour de la droite pour de faibles valeurs de x alors qu'ils s'en écartent davantage pour de grandes valeurs. I

Il est vraisemblable que **la supposition d'une variance identique des résidus n'est pas vérifiée**.

X(E)-Y(E)



La droite est obtenue uniquement à cause du point suspect à droite. Les autres points étant alignés verticalement, la droite de régression ne peut être calculée.

La droite obtenue ici ne correspond à rien.

Il est essentiel de **toujours** regarder le nuage de points **<u>avant toute chose</u>**, pour s'assurer de la validité de la démarche.

5. Interprétation des résultats

Une fois les hypothèses vérifiées graphiquement, les coefficients de corrélation et détermination calculés vont nous permettre d'interpréter les résultats… en restant prudent quant au lien de cause à effet.

- ◊ le coefficient de corrélation nous donne des informations sur l'existence d'une relation affine entre les deux grandeurs considérées ;
- ◊ il ne faut pas confondre corrélation et relation causale : une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.
- ◊ l'existence d'une corrélation n'est jamais la preuve d'une relation de cause à effet.
- ◊ l'ajustement parfait correspond à $r^2 = 1$;
- ◊ une valeur élevée de r (très proche de -1 ou de +1) ne ment jamais ;
- ◊ l'ajustement est d'autant meilleur que $|r|$ est élevé ;
- ◊ l'ajustement est d'autant plus mauvais que r se rapproche de 0 ;
- ◊ une valeur faible (très proche de 0) du coefficient de corrélation ne permet pas de tirer de conclusion car elle peut recouvrir des réalités trop différentes ;
- ◊ un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux grandeurs car il peut exister une relation non linéaire entres elles.
- ◊ si les variables sont indépendantes alors la corrélation est nulle ;
- ◊ si les variables ne sont pas indépendantes, le nuage prend la forme d'une ellipse d'autant plus aplatie que la corrélation est forte ;
- ◊ le coefficient de corrélation est très sensible aux données aberrantes ;
- ◊ le coefficient de corrélation ne dépend pas des unités dans lesquelles sont exprimées les observations.
- ◊ Le coefficient de détermination détermine la proportion (ou pourcentage) de la variation (ou variance) de la variable Y imputable à la variable X.