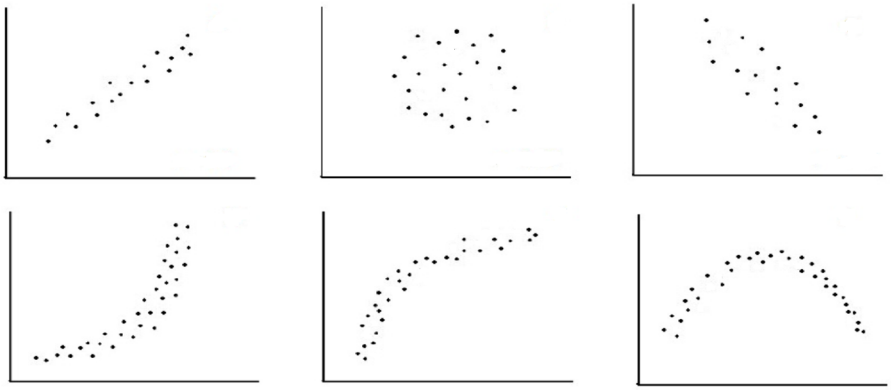
La fiche élève

La droite de régression linéaire, ou droite d'ajustement linéaire

Régression : « Réduction de données complexes, prélevées par lots sur un phénomène physique ou économique, à une donnée plus simple qui fait parfois apparaître une loi cachée. » (*)

Problématique :

Lors de la représentation de séries statistiques doubles $(x_i\,;\,y_i)$ on constate qu'il existe des nuages de points ayant des formes remarquables pouvant rappeler des courbes connues :



Parmi ceux-ci, il y a des nuages dont les points respectent un alignement approximatif, comme le premier et le troisième (forme ovale).

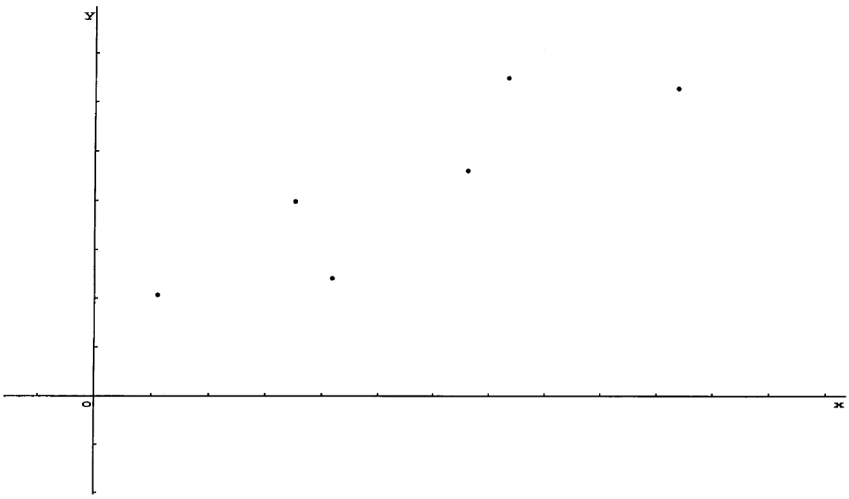
Le but de la séance est de déterminer alors le meilleur *modèle* mathématique possible sous la forme d'une fonction affine $f(x) = y = ax + b$ qui lie les variables y et x. Pour que ce modèle théorique décrive le mieux possible le problème réel, il faut que sa droite représentative « approche le mieux possible » le nuage de points. On donnera un sens précis à cette proposition.

(*) Le mot « **Régression** » désignant la droite d'ajustement linéaire est bien étrange ; il est dû à un cousin de Charles Darwin, Sir Francis Galton (1822-1911), et à son étude sur la taille des enfants.
Il avait observé un phénomène, bien connu depuis, que celle-ci tend toujours à se rapprocher de la taille moyenne ; F. Galton, scientifique dont les théories ont été souvent contestées, a donc parlé de RÉGRESSION au sens littéral : « recul, amoindrissement, diminution ».

L'ensemble des couples $(x_i ; y_i)$ donne des résultats observés sur la population.	La fonction affine $f(x) = y = ax + b$ **donne un modèle théorique** pour représenter la situation.
Observé	Réel

Approche du problème :
Pour chaque partie qui va suivre, vous êtes libres de faire autant d'essais qu'il vous semble utile pour en tirer des conclusions que vous noterez avec soin et de manière la plus concise possible (on utilise toutes les finesses du français !).

Avec la méthode de votre choix (graphique, distance aux points, autre, …), construire une droite qui « approche bien » le nuage de points (c'est-à-dire telle que les points se trouvent les plus « proches » possibles de la droite).



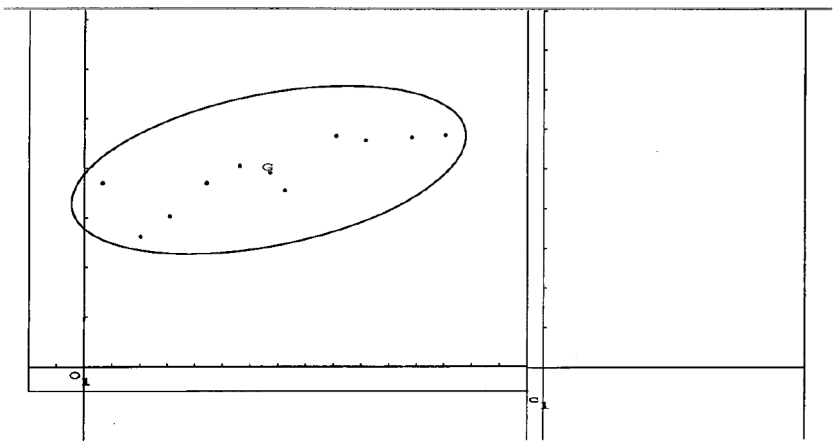
Justifier votre choix :

Confronter alors ce choix avec d'autres. Quelle méthode retenir au final ?

Partie A : À la découverte de la droite.

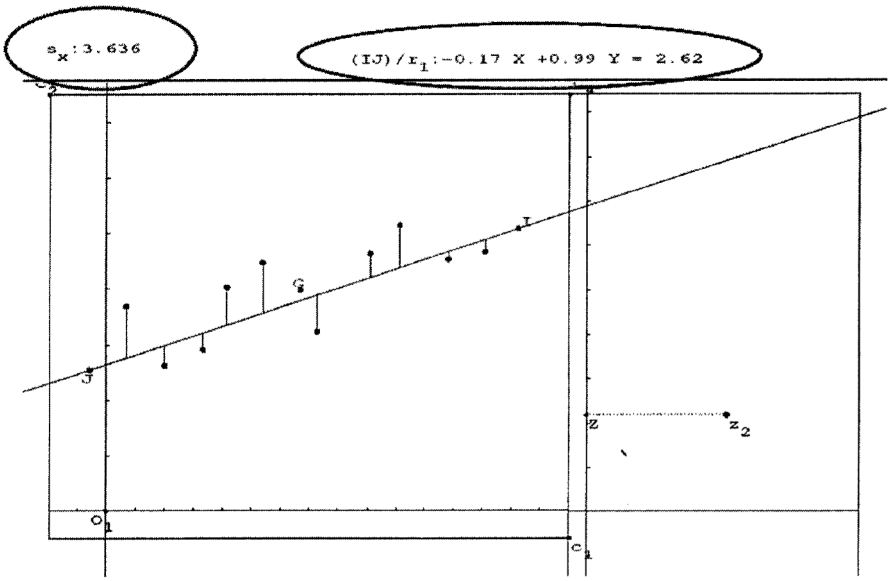
a) Ouvrir « Reg0.G2w ». Ce fichier Géoplan contient un **nuage de points** représentant une **série statistique double** x_i en abscisse et y_i en ordonnée. Chaque point du nuage est déplaçable à volonté en cliquant sur celui-ci avec le clic gauche et en le maintenant pendant le transport. G est le point moyen. Ce fichier comprend aussi une droite définie par deux points I et J qui sont eux aussi déplaçables avec la souris.

b) Constituer un nuage de points répartis pour différentes valeurs de x et de y de « forme ovale », par exemple :



Les points du nuage sont maintenant fixés pour le reste de l'exercice.

c) On veut définir avec précision ce que veut dire « approcher le nuage » pour la droite. Pour cela, la touche X du clavier peut faire apparaître (ou disparaître) pour chaque point du nuage la distance de ce point au point de la droite de même abscisse. Cette distance est appelée **résidu**, c'est-à-dire le « restant ».



On considère alors la somme des carrés de ces distances que l'on peut écrire : $S_x = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2$. Cette somme, appelée **somme des carrés des résidus**, apparaît en dynamique en haut et à gauche de l'écran en valeur approchée. Le point Z_2 dans le repère de droite a une ordonnée proportionnelle à S_x, pour tous les fichiers de cette activité.

d) En ne déplaçant que les points I et J, il faut trouver, par essais successifs, une droite D qui minimise la somme S_x. Une fois celle-ci déterminée, noter la valeur de la somme correspondante, ainsi que l'équation cartésienne de la droite et la position de G par rapport à D. Donner l'équation cartésienne réduite de D.

Recommencer l'opération avec un nuage de points de « direction très différente ». Noter de même les données obtenues.

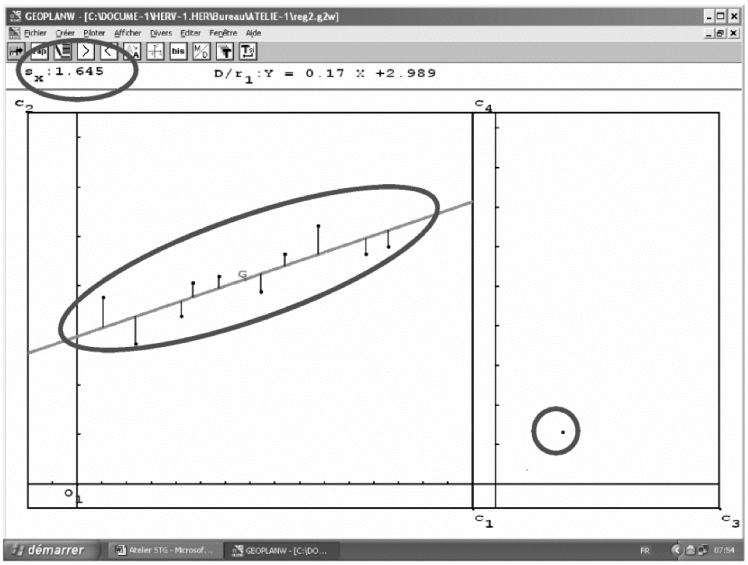
Partie B : La droite et le point moyen.

Ouvrir « Reg1.G2w ». Disposer les points du nuage.
Les flèches « Haut » et « Bas » permettent de déplacer la droite D parallèlement, c'est-à-dire à **coefficient directeur constant**.
La touche trace « T » permet de garder la trace du point Z_2, dont l'abscisse augmente avec la distance verticale de G au point de la droite D de même abscisse.
Déplacer alors la droite pour déterminer graphiquement celle qui minimise S_x.
Que constate-t-on pour le point G ? Noter les données obtenues.
Faire varier la forme du nuage et recommencer l'opération. Ce qui a été vérifié pour G semble-t-il confirmé ? Noter dans le cadre ci-dessous cette observation.

Première conclusion :

Partie C : La bonne direction.

Ouvrir « Reg2.G2w ». Maintenant que nous avons déterminé un critère de position à coefficient directeur constant, nous allons déterminer, parmi toutes les droites qui passent par G, s'il en existe une qui minimise la somme des carrés des résidus, avec la « bonne direction ».



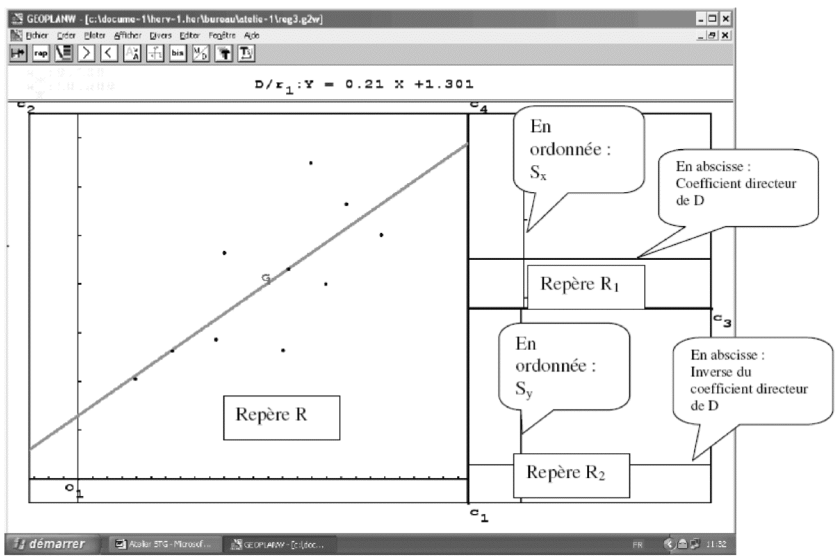
Le point Z_2 dans le repère R_1 situé à droite a pour abscisse la valeur absolue du coefficient directeur de D et pour ordonnée la valeur de S_x. D se déplace avec les flèches « haut » et « bas ».

Quelle est la position de la droite D lorsque Z_2 est sur l'axe des ordonnées ?
Déterminer graphiquement la droite qui minimise S_x et noter les données approchées obtenues.
Reprendre cette observation avec un autre nuage, puis éventuellement un troisième si nécessaire.
Quelle(s) conclusion(s) peut-on tirer des parties B et C de cette étude ?

Conclusion :

Partie D : Un autre choix de résidus.

Ouvrir « Reg3.G2w ».
Le choix a été fait de mesurer les distances à la droite verticalement. On peut refaire cette étude en prenant les distances à la droite mesurée horizontalement. De manière analogue à la partie A, on considère la somme S_y des carrés de ces distances.
On admet que le point G appartient aussi à la droite qui minimise S_y.
Ces sommes sont représentées respectivement dans les repères situés en haut à droite et en bas à droite, comme expliqué dans la figure suivante.



La touche « Y » permet de faire afficher ces distances. Déterminer alors à l'aide des flèches la droite qui minimise la somme des carrés des résidus affichée en haut S_y et en utilisant la touche « X » la droite qui minimise S_x (nommée *droite d'ajustement* de *y* en *x* ou *droite de régression linéaire y* en *x*).

Est-ce la même droite ? Noter les équations respectives affichées, ainsi que les valeurs de S_x et S_y.

Les droites qui minimisent S_x et S_y peuvent être affichées avec la touche « W » (*droite d'ajustement* de y en x qui donne $y = f(x)$) ou la touche U (*droite d'ajustement* de x en y qui donne $x = g(y)$).
En déplaçant les points du nuage, peut-on rendre confondues les deux droites ? Si oui, dans quel cas ?

Conclusion(s) de la partie D :

Partie E : Des essais de modification.

Ouvrir « Reg4.G2w ». Le but de cette partie est d'observer l'effet sur la droite de régression du déplacement d'un, ou de plusieurs points du nuage.
Déplacer des points et observer l'effet sur le point moyen.
La touche « B » fait apparaître la droite de régression de y en x. Observer sa variation lors du déplacement de points, notamment lors du déplacement de points sur la droite.

Pour compléter ces observations, la touche « M» fait apparaître un nouveau point du nuage déplaçable, le nouveau point moyen G', la nouvelle droite de régression du nouveau nuage.

Observations de la partie E :

Partie F

Quel résumé peut-on faire de toute cette étude ?