Une activité en classe terminale STG : La droite de régression linéaire(*)

Hervé Milliard(**)

1. Introduction

Dans le cadre des nouveaux programmes de la série STG, nous avons voulu mettre en place une activité qui réponde à de multiples critères :

- Se conformer aux objectifs généraux de la série,
- Traiter d'un point précis et important du programme,
- Rendre accessible aux élèves de cette série une notion théoriquement difficile à leur niveau,
- Mobiliser des capacités de lecture, d'observation, d'expérimentation, d'analyse et de synthèse,
- Apporter une petite contribution d'histoire des maths,
- Montrer la force des nouvelles technologies pour réaliser ce pari difficile et l'économie de temps qu'elles permettent, la possibilité de reprendre, de tester des cas différents, d'agir en modifiant un seul facteur ou plusieurs simultanément etc.

2. Les objectifs généraux de la série (programme de la série STG. 2005)

Les objectifs suivants sont prioritairement visés :
– entraîner à la lecture active de l'information, à sa critique, à son traitement, en particulier en privilégiant les connaissances et les méthodes permettant des changements de registre (graphique, numérique, algébrique, …) ;
– former les élèves à l'activité scientifique par l'acquisition de méthodes d'observation, d'analyse critique et de déduction ;
– développer les capacités de communication écrite et orale;
– promouvoir la cohérence de la formation des élèves en utilisant les liens entre les différentes parties du programme et en tissant les relations entre les mathématiques et les autres disciplines.

3. Le programme (extraits du programme concernant le thème)

Le programme de statistique est un terrain pour des activités interdisciplinaires et pour la consolidation des techniques élémentaires de calcul : usage des fractions, des pourcentages, proportionnalité. Les statistiques à deux variables sont indispensables en économie et en gestion pour analyser, interpréter et prévoir.

(*) Tous les fichiers utilisés dans ce document sont téléchargeables sur le site académique de l'académie d'Aix-Marseille à l'adresse suivante :
http://www.maths.ac-aix-marseille.fr/tic/classe/lgt/ajust-affine/

(**) Professeur au lycée de Marseille-Veyre.

Contenus	Capacités attendues	Commentaires
Étude de séries de données statistiques quantitatives à deux variables		On accompagne ce travail d'un entretien des capacités sur les statistiques à une variable de la classe de première.
Nuage de points, point moyen.	Associer un tableau de données à la suite (x_k,y_k), $0 < k < N + 1$, où N est l'effectif de la population.	
	Représenter graphiquement un nuage de points et déterminer le point moyen.	Le point moyen a pour coordonnées $(\overline{x},\overline{y})$.
Ajustement affine.	Trouver une fonction affine qui exprime de façon approchée y en fonction de x. Utiliser cette fonction pour interpoler ou extrapoler.	L'objectif est d'étudier le lien éventuel entre deux caractères d'une même population.
Séries chronologiques.	Utiliser un ajustement affine pour faire une prévision.	L'ajustement est réalisé soit par une méthode graphique, soit par la méthode des moindres carrés à l'aide de la calculatrice ou du tableur.

4. Les commentaires du programme (doc. Accompagnement. p. 17)

En terminale, sont introduites les séries de données statistiques à deux variables. L'objectif est d'étudier le lien éventuel entre deux caractères d'une même population. Pour deux variables quantitatives, on peut rechercher une formule approchée exprimant l'une des variables en fonction de l'autre : ***c'est la problématique de l'ajustement****.*
En première approximation, l'ajustement affine peut être réalisé ***graphiquement*** *; il s'agit alors de tracer, « à la main », quand la forme du nuage de points s'y prête, une droite passant « au plus près » des points du nuage.*

Si l'objectif fixé par le problème le justifie, la calculatrice ou le tableur permettent un ajustement par la méthode des moindres carrés. On peut néanmoins en expliquer le principe :
Pour une droite donnée d'équation $y = ax + b$, on compare les y_i observés aux y_i calculés (autrement dit les $ax_i + b$) en calculant les résidus $y_i - (ax_i + b)$. On souhaite trouver une droite pour laquelle ces résidus soient les plus faibles possibles. La droite des moindres carrés, ou droite de régression, est celle qui minimise la somme des carrés de ces résidus.

Pour un nuage donné, il est intéressant de comparer, à l'aide du tableur, la somme des carrés des résidus pour plusieurs droites obtenues par différentes façons (graphiquement, calculatrice, tableur) ; on vérifie, à cette occasion, que la droite de régression fournie par le tableur, donne une somme des carrés des résidus plus faible que les autres.

On veillera à ne traiter que des problèmes où l'ajustement affine a un sens (forme du nuage de points presque rectiligne).

On pourra traiter un ou deux exemples de situations concrètes où les deux ajustements ont un sens et permettent de mettre en évidence la non-symétrie du rôle joué par chacune des deux variables.
Le professeur peut faire remarquer aux élèves que la droite de régression de y en x n'est pas la même que celle de x en y, que ces deux droites ne sont donc pas interchangeables (sauf dans le cas extrême où les points sont alignés) et veiller à ne pas demander une estimation de x en y en utilisant la droite de régression de y en x.

5. Les objectifs de la séance

La trame de l'activité proposée – qui peut se faire en demi groupe en salle informatique ou en classe entière avec un vidéoprojecteur – après avoir présenté avec soin la problématique, reprend l'idée de la démonstration théorique de la droite de régression. On montre d'abord qu'à coefficient directeur constant, la somme des carrés des résidus S_x est minimale lorsque la droite passe par le point moyen, puis que, de toutes les droites passant par le point moyen, il y en une seule qui minimise S_x.
Lorsque cette notion est acquise, on reprend l'étude de manière analogue de la droite qui minimise S_y et on permettra enfin aux élèves de « jouer » avec ces connaissances en modifiant à volonté les nuages de points pour observer les variations sur S_x, S_y, et les droites d'ajustement.
L'objectif est finalement de faire comprendre ce qu'est la droite de régression, et le fait qu'elle est unique (existence et unicité) et non d'insister sur le calcul des coefficients de cette droite.
On pourra dire en fin de séance que l'expression $y = ax + b$ est donnée par la calculatrice avec une excellente précision, suffisante pour la plupart des applications et la donner sous forme $y = a(x - \overline{x}) + \overline{y}$.

Privilégiant là une démarche scientifique relativement autonome, une grande liberté est donnée aux élèves pour effectuer leur recherche, faire des essais complémentaires, rédiger les conclusions de chaque partie.
Le professeur est présent et prêt à répondre à toutes les questions et toutes les petites angoisses. Il fera notamment la distinction entre l'exact et l'approché tout au long de la séance.
Cette activité, faut-il le préciser, se fait après le cours sur les séries statistiques doubles.
Cette séance sera de toutes façons suivie de bien d'autres qui permettront de montrer tout l'intérêt de cette droite pour attribuer des valeurs de séries, pour effectuer des prévisions, etc.

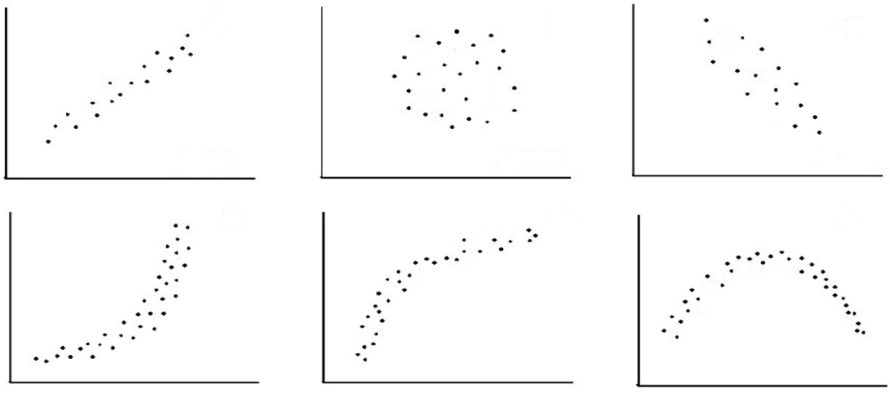
La fiche élève

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.

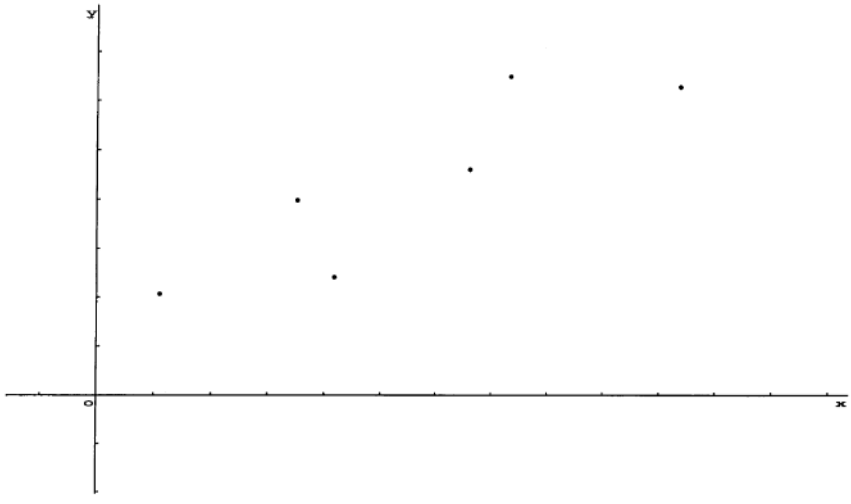
Observé

La fonction affine $f(x) = y = ax + b$ **donne un modèle théorique** pour représenter la situation.

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).



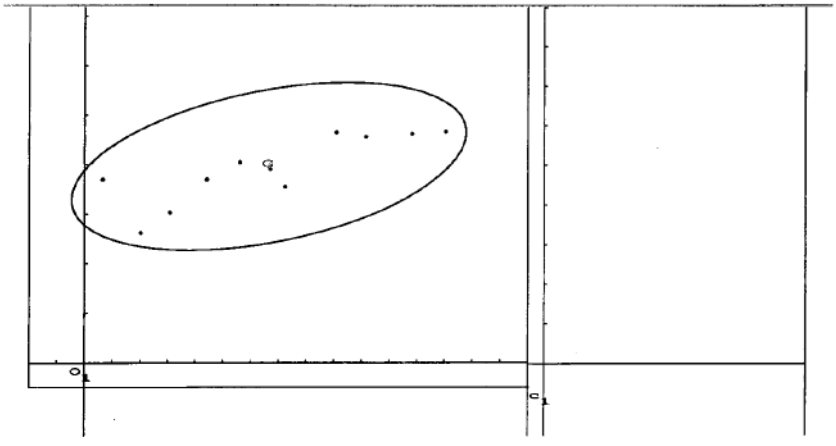
Justifier votre choix :

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

Partie A : À la découverte de la droite.

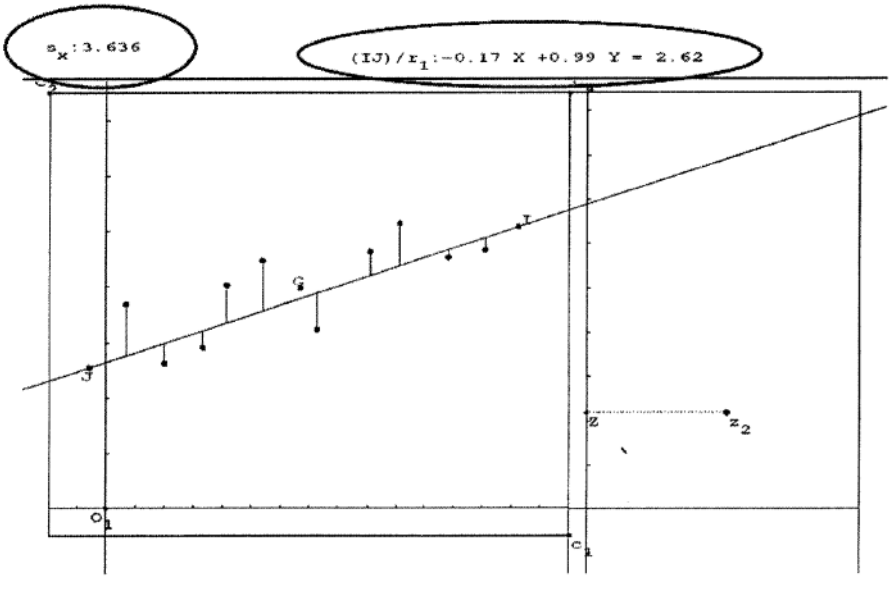
a) Ouvrir « Reg0.G2w ». Ce fichier Géoplan contient un **nuage de points** représentant une **série statistique double** x_i en abscisse et y_i en ordonnée. Chaque point du nuage est déplaçable à volonté en cliquant sur celui-ci avec le clic gauche et en le maintenant pendant le transport. G est le point 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.

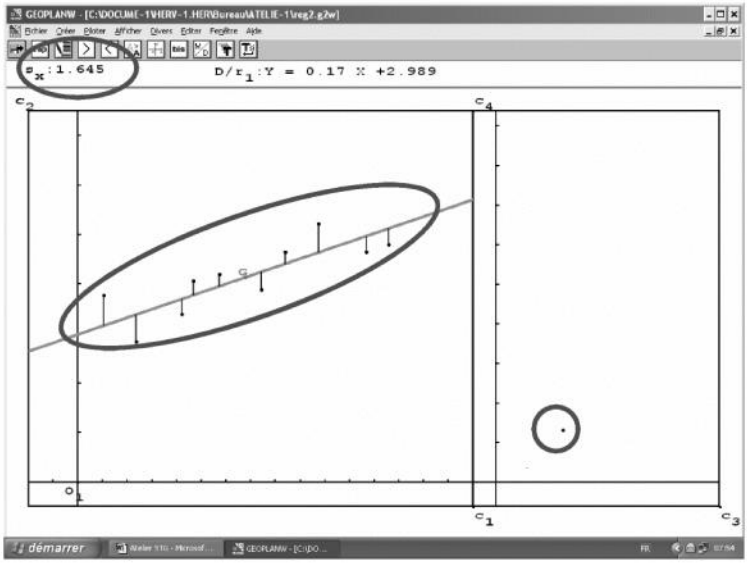
Partie B : La droite et le point moyen.

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 ».



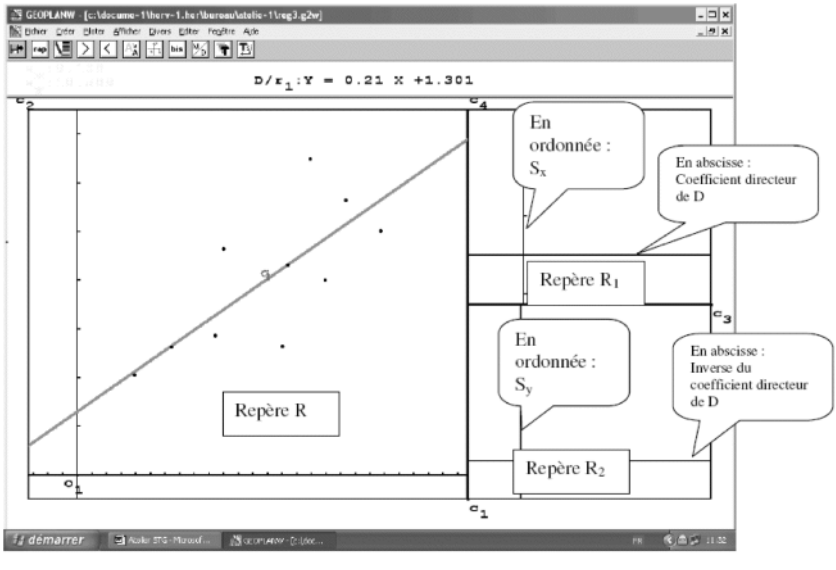
Le point Z_2 dans le repère R_1 situé à droite a pour abscisse la valeur absolue du coefficient directeur de D et pour ordonnée la valeur de S_x. D se déplace avec les flèches « haut » et « bas ».

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 ?

Compte rendu de séance :

La séance a lieu en une heure 15 min en salle informatique avec deux élèves par poste. Les élèves disposent des fiches et des fichiers Géoplan sur l'ordinateur, sans document papier.
Ils ont déjà utilisé le logiciel trois fois, ce qui leur a donné un peu de pratique et permis de commencer à travailler l'activité dès le début. Le professeur a comme ligne de conduite d'intervenir le moins possible, la fiche étant très détaillée.
Ce choix de détail a été délibérément fait pour éviter des questions récurrentes d'ordre pratique et permettre une réelle activité d'enseignement et de recherche en se concentrant sur les deux questions suivantes :

- Peut-on donner un sens précis en définissant un réel qui indique si une droite passe « plus ou moins près » des points du nuage.
- Existe-t-il alors une droite *meilleure que toutes les autres* au sens défini ?

On pourra en annexe comparer les droites d'ajustement de y en x et de x en y et observer la variation de ces droites lorsqu'on déplace des points ou que l'on en rajoute.

Cette activité de la partie A répond au programme officiel :
« *En première approximation, l'ajustement affine peut être réalisé* ***graphiquement*** *; il s'agit alors de tracer " à la main ", quand la forme du nuage de points s'y prête, une droite passant " au plus près " des points du nuage.* »
Il est enfin utile de rappeler que cette étude reprend graphiquement l'esprit de la démonstration qui mène à la détermination de la droite de régression.

Partie A

Les élèves ont dû prendre un peu de temps pour comprendre la problématique et prendre en main le fichier. Ils ont tout de suite compris que G représente le point moyen et qu'il est donc logique qu'il se déplace avec les points du nuage.
La somme S_x et surtout son expression ont suscité plusieurs questions. Le professeur a dû faire préciser au tableau avec un graphique ce qu'elle mesure précisément.
Les élèves, au final de cette partie, trouvent des résultats sensiblement différents, le fait de jouer sur les deux variables (direction, ordonnée à l'origine) ne leur ayant pas toujours permis de déterminer la droite de régression (bien qu'assez proche).
Ceci n'est pas gênant (au contraire !) car l'essentiel de cette partie était de comprendre que l'on peut créer un nombre positif qui mesure la « *distance de la droite au nuage* ».
La méthode qui a remporté le plus vif succès est une méthode purement graphique, en déplaçant la droite jusqu'à ce que sa position par rapport au nuage soit « satisfaisante » pour l'œil.

Partie B

Les élèves découvrent très rapidement que la droite qui minimise S_x passe par G, un deuxième essai avec un nuage très différent confirmant immédiatement leur conjecture.
J'ai pu constater une nouvelle fois la faculté chaque année meilleure des élèves de s'adapter rapidement à l'outil informatique.

Partie C

Si les élèves n'ont eu aucune difficulté avec la détermination de la droite de régression, ils ont eu plus de mal à comprendre l'abscisse du point Z_2 dans le repère R_2 :
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. Ceci a été l'occasion de plusieurs rappels sur la représentation graphique, la valeur absolue, le coefficient directeur d'une droite et son rôle graphique.
La première conclusion (en fait l'essentiel de la recherche) a mené globalement à des réponses correctes, parfois inexactes ou incomplètes : « la droite idéale est la droite qui passe par G » ; « Il y a deux droites qui minimisent S_x : l'une qui passe par G, l'autre de coefficient directeur donné »

Partie D

À l'aide des repères R_2 et R_3, les élèves ont rapidement compris que ce ne sont pas les mêmes droites qui minimisent S_x et S_y. Certains ont modifié spontanément les points du nuage pour observer l'effet sur les droites de régression.
Ceci a permis à la plupart de comprendre que c'est en alignant les points que l'on rend confondues ces droites.
On est très proche des consignes du programme :
Le professeur peut faire remarquer aux élèves que la droite de régression de y en x n'est pas la même que celle de x en y, que ces deux droites ne sont donc pas interchangeables (sauf dans le cas extrême où les points sont alignés) et veiller à ne pas demander une estimation de x en y en utilisant la droite de régression de y en x.
Dans ce cas, ce sont les élèves eux-mêmes qui ont pu faire ces observations.

Partie E

Cette partie est beaucoup plus ouverte. Elle laisse les élèves manipuler sans but précis affiché si ce n'est d'observer comment « bouge » la droite lorsqu'on déplace les points du nuage, soit près de G, ou au contraire éloignés, parallèlement à D ou pas, sur la droite elle-même…
L'intérêt pédagogique est multiple : placer l'élève en situation de recherche pour dégager tout ce qui peut apporter des résultats intéressants, voir par exemple qu'introduire un nouveau résultat dont le point correspondant est sur D ne modifie pas une prévision.

Quelques extraits :
« La droite varie beaucoup lorsqu'on déplace un seul point loin de D »
« Pas de changement si on bouge un point sur la droite »
« Si on rajoute un point très éloigné, la droite change parce que G change »
« Quand on rajoute un point au nuage, mais sur D, la droite ne change pas »

L'activité s'est achevée en une heure quinze, en prenant tout le temps nécessaire. Le professeur a eu soin de demander, pour chaque partie, plusieurs réponses écrites d'élèves et de faire une mise en commun qui permette d'établir au final la solution correcte.
Le résumé de cette étude destiné au cours est établi au cours de l'heure suivante.

En conclusion

La démonstration qui mène à la droite d'ajustement n'est pas au programme des séries STG ou ES, mais ici les nouvelles technologies vont permettre d'en reprendre l'idée en montrant graphiquement dans un premier temps que, à coefficient directeur fixé, la somme des carrés des résidus est minimale lorsque la droite passe par le point moyen, puis dans un deuxième temps que, pour toutes les droites qui passent par G, S_x est minimale lorsque la droite a un coefficient directeur bien déterminé.
Le professeur pourra demander aux élèves d'utiliser le logiciel pour montrer, questionner, faire conjecturer et faire confirmer par plusieurs essais avant de formaliser.

L'expérience réalisée en classe a montré que les élèves mènent en autonomie l'essentiel de l'activité, mais que le professeur reste précieux pour éclaircir, pour préciser, pour formuler correctement. Ce travail, par l'abondance des essais possibles, par la possibilité de « faire bouger » par translation ou par rotation, par la visualisation sur un graphique associé, montre la force des nouveaux outils pour assimiler une notion difficile.
La suite a montré que la notion a été bien assimilée et que l'on a pu passer à l'utilisation de la droite de régression pour résoudre des problèmes pris dans les différentes disciplines.

Bibliographie de complément sur le sujet :

http://www.unilim.fr/pages_perso/jean.debord/math/reglin/reglin.htm
http://www.geog.umontreal.ca/donnees/geo1522/Cours3.ppt
http://www.cepremap.cnrs.fr/~michel/metri2/cours_1.pdf
Et un livre : *Comment interpréter les résultats d'une régression linéaire ?* de R. Tomassone.

L'activité présentée dans cet article a été proposée par le groupe de recherche académique lycée d'Aix-Marseille, lors de journées inter académiques en décembre 2005 à Lyon.
La mise au point de l'article doit beaucoup aux animateurs de cet atelier, et particulièrement à Jean-Pierre Sicre et Stéphane Clément, que je tiens à remercier.