La droite de régression enfin révélée

André Boileau
Professeur retraité, UQAM
GRMS, octobre 2016

Résumé

Je me suis longtemps demandé comment aborder la régression linéaire de façon à la rendre compréhensible pour des élèves de niveau secondaire, et je crois enfin y être parvenu.

Cet atelier vise à révéler tout ce que vous avez toujours voulu savoir sur la droite de régression sans jamais oser le demander. Fini les explications compliquées ou vaseuses, bienvenue à la simplicité et à la clarté, pour ne pas dire à l'évidence.

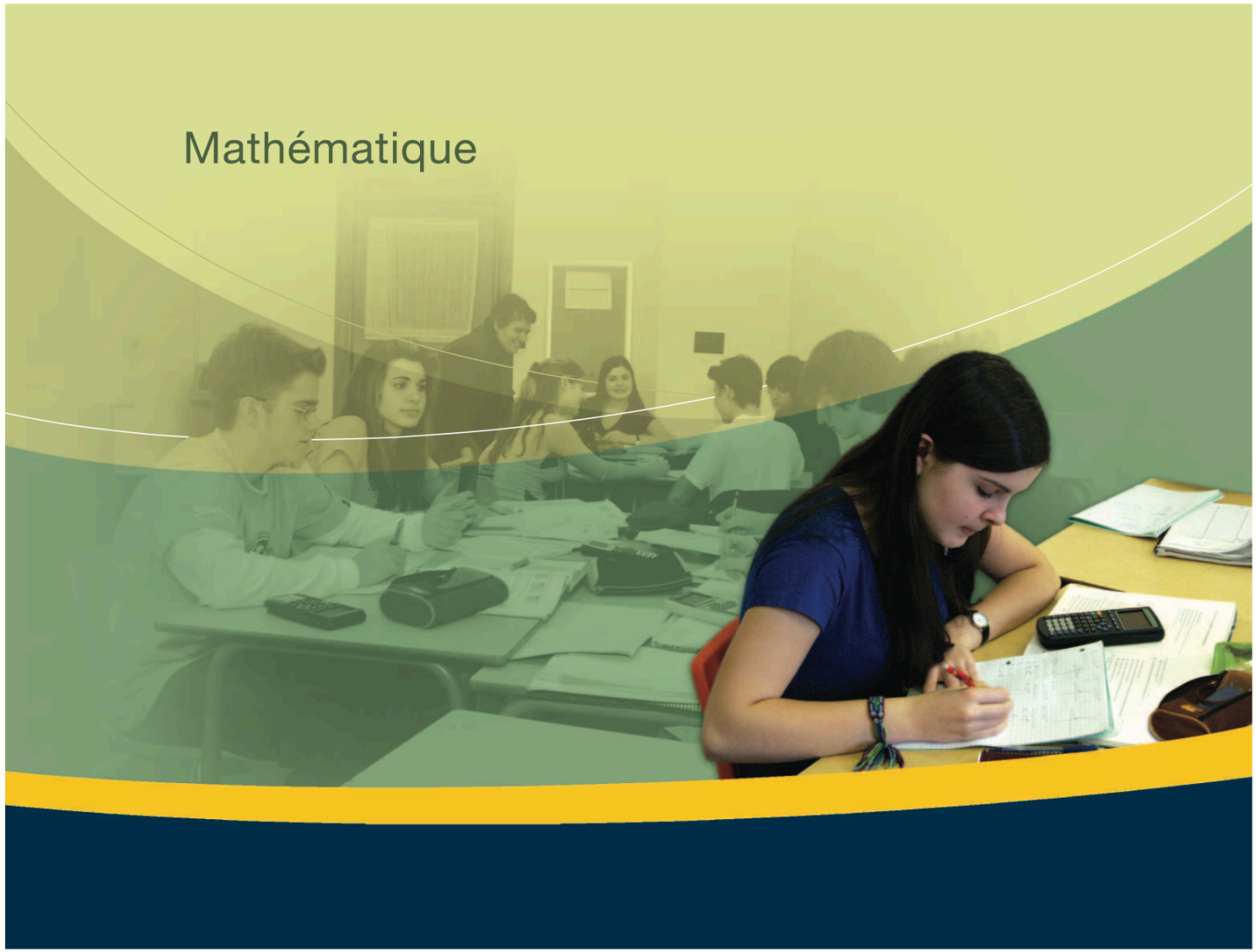
Amateurs de mathématiques formelles, prière de s'abstenir, car vous serez déçus…

Apportez votre ordinateur portatif avec GeoGebra installé… ça pourra toujours servir!

Plan de l'atelier

- Pourquoi parler de régression (linéaire) au secondaire ?
- Comment parle-t-on de régression linéaire au secondaire ? Contenu mathématique de ces discours...
- Doit-on définir précisément la régression linéaire ?
- Une approche utilisant la technologie
  - Aide à définir et à comprendre le pourquoi de cette définition
  - Aide à trouver approximativement la droite de régression tout en préservant les intuitions sous-jacentes
- Conclusion / Questions / Discussions

Programme de formation de l'école québécoise



Deuxième année du cycle

Dans cette séquence, l'élève poursuit son apprentissage de la statistique descriptive et apprend à faire intuitivement quelques inférences. À l'instar des probabilités, la statistique est un outil qui contribue à la prise de décisions. Pour répondre à des questions d'ordre pratique ou social, l'élève recueille des données, les organise, les représente et détermine différentes mesures. Il choisit le diagramme qui convient le mieux à la distribution et aux informations qu'il veut représenter. Il vérifie si la distribution contient des données aberrantes (ou extrêmes) susceptibles d'influencer certaines mesures et ses conclusions. Il est aussi attentif aux biais qui pourraient, tout au long du processus, nuire à la fiabilité de l'étude. Il les relève et les corrige, s'il y a lieu. Pour analyser et comparer des distributions, il observe leur forme et utilise les mesures de tendance centrale et de dispersion appropriées[26]. Il dégage les avantages et les limites des différentes mesures de dispersion : étendue, étendue interquartile, écart moyen.

Précédemment, pour développer le sens des liens de dépendance, l'élève a entrepris de façon implicite et expérimentale l'étude de distributions statistiques à deux caractères par l'introduction du nuage de points et de l'estimation de la droite de régression. L'analyse du nuage de points permet non seulement de se renseigner sur la corrélation entre les variables, mais aussi de la caractériser. Ce diagramme est représenté à partir d'un tableau de distribution à deux caractères présentant les résultats tirés d'un relevé fourni ou d'une expérience réalisée. L'élève apprend à interpréter qualitativement la corrélation : positive ou négative, nulle, forte ou faible, parfaite ou imparfaite. Il évalue approximativement le coefficient de corrélation à l'aide d'une méthode graphique et calcule, au besoin, sa valeur à l'aide de la technologie. L'analyse et l'interprétation des situations sont importantes. L'élève est sensibilisé au fait que même si la corrélation est forte, cela ne signifie pas nécessairement qu'il existe un lien de causalité. En effet, cette relation apparente peut être aussi fortuite ou liée à un troisième facteur.

Dans le cas de la corrélation linéaire, l'élève trace la droite la mieux ajustée en tenant compte des données aberrantes (ou extrêmes), détermine la règle de cette droite et fait des extrapolations. Il est amené à prendre conscience que la fiabilité de l'interpolation ou de l'extrapolation dépend du degré de dépendance entre les deux caractères. Pour tracer cette droite ou déterminer son équation, il peut réinvestir les concepts de médiane ou de moyenne selon qu'il effectue son ajustement à l'aide des méthodes de la droite médiane-médiane ou de la droite de Mayer[27]. Il peut aussi utiliser la technologie. L'étude de la méthode des moindres carrés n'est toutefois pas au programme de cette séquence.

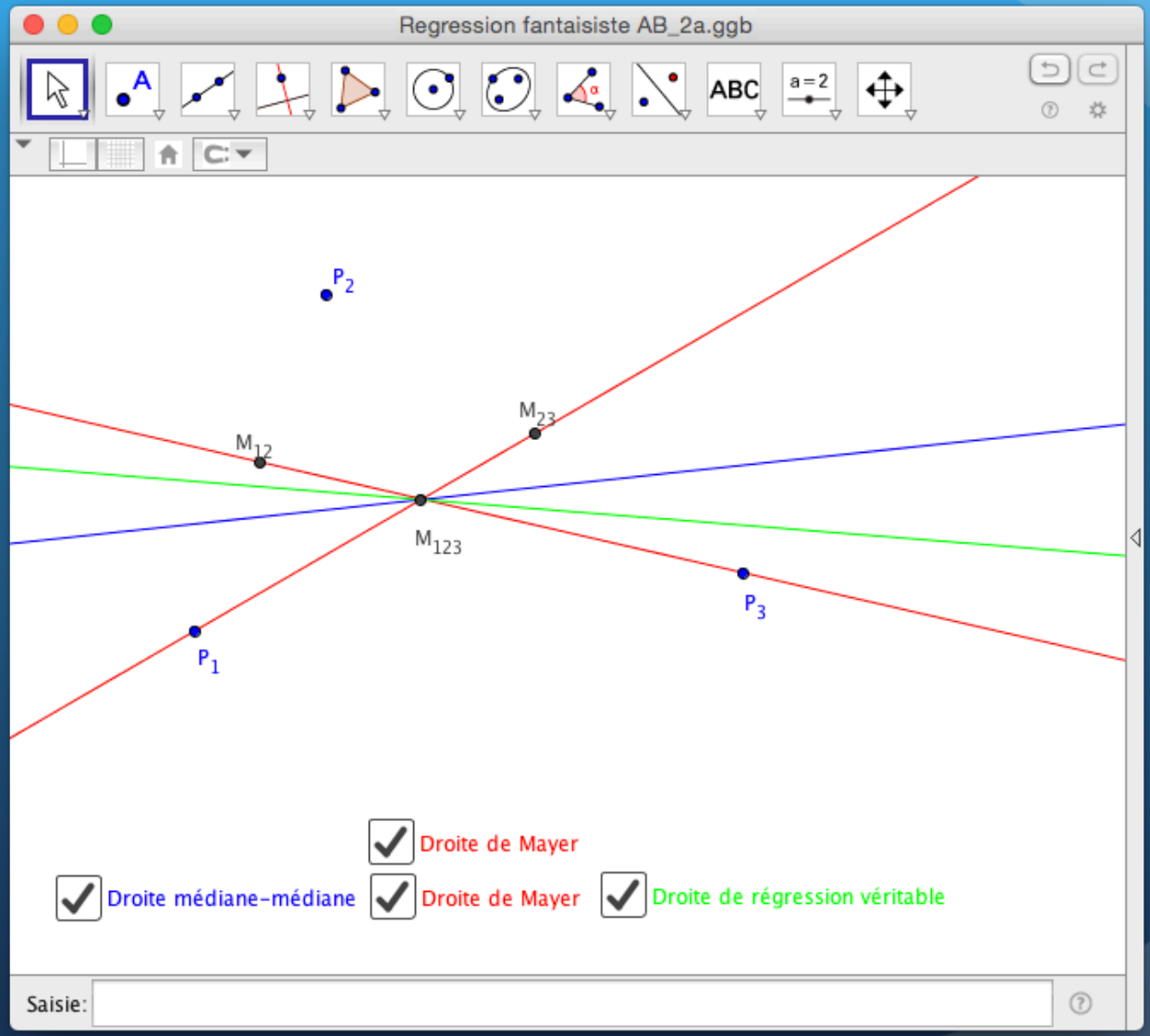
Troisième année du cycle

À la dernière année du cycle, l'élève réinvestit ses savoirs liés à la statistique dans différentes situations et plus particulièrement dans la réalisation de son activité synthèse.

26. Pour caractériser une distribution, on utilise habituellement deux mesures statistiques : une mesure de tendance centrale et une mesure de dispersion. Le couple le plus utilisé est celui formé de la moyenne et de l'écart type, mais on utilise aussi la médiane et l'intervalle interquartile dans le cas où la distribution est asymétrique. Dans la séquence *Culture, société et technique*, l'écart type n'est pas au programme. L'élève analyse la distribution à l'aide de l'écart moyen.
27. La méthode de la droite médiane-médiane consiste à partager les données en trois groupes (les premier et troisième groupes doivent avoir le même nombre de données), à calculer les médianes de chacun des groupes et à tracer la droite qui passe par le point moyen de ces trois médianes et qui est parallèle à la droite qui passe par les première et troisième médianes. Pour construire la droite de Mayer, on partage les données en deux groupes et on calcule les coordonnées des points moyens pour chaque groupe. On trace la droite qui passe par ces deux points.

27. La méthode de la droite médiane-médiane consiste à partager les données en trois groupes (les premier et troisième groupes doivent avoir le même nombre de données), à calculer les médianes de chacun des groupes et à tracer la droite qui passe par le point moyen de ces trois médianes et qui est parallèle à la droite qui passe par les première et troisième médianes. Pour construire la droite de Mayer, on partage les données en deux groupes et on calcule les coordonnées des points moyens pour chaque groupe. On trace la droite qui passe par ces deux points.

Approche pseudo-mathématique

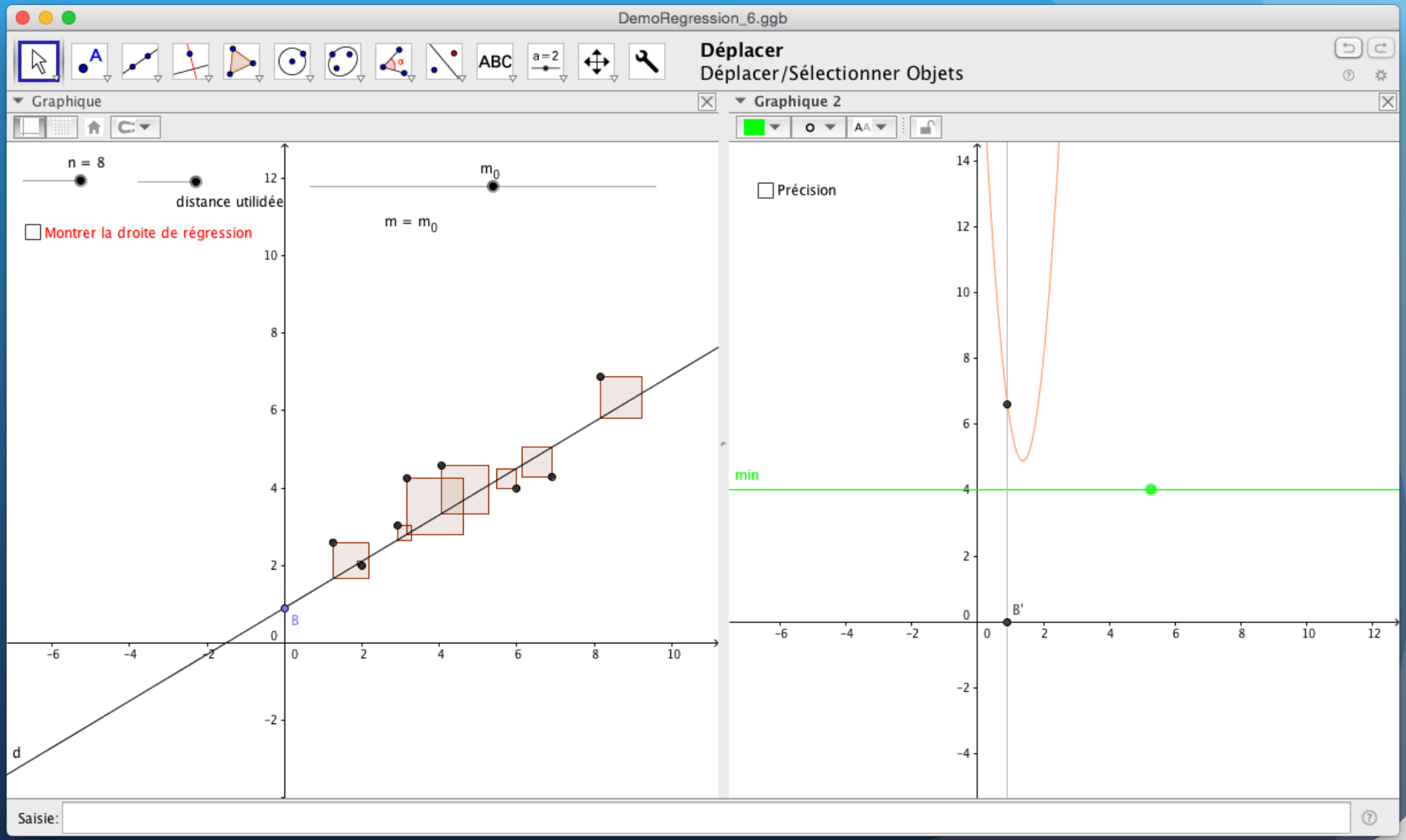


Approche mathématique précise ?

DÉFINITION

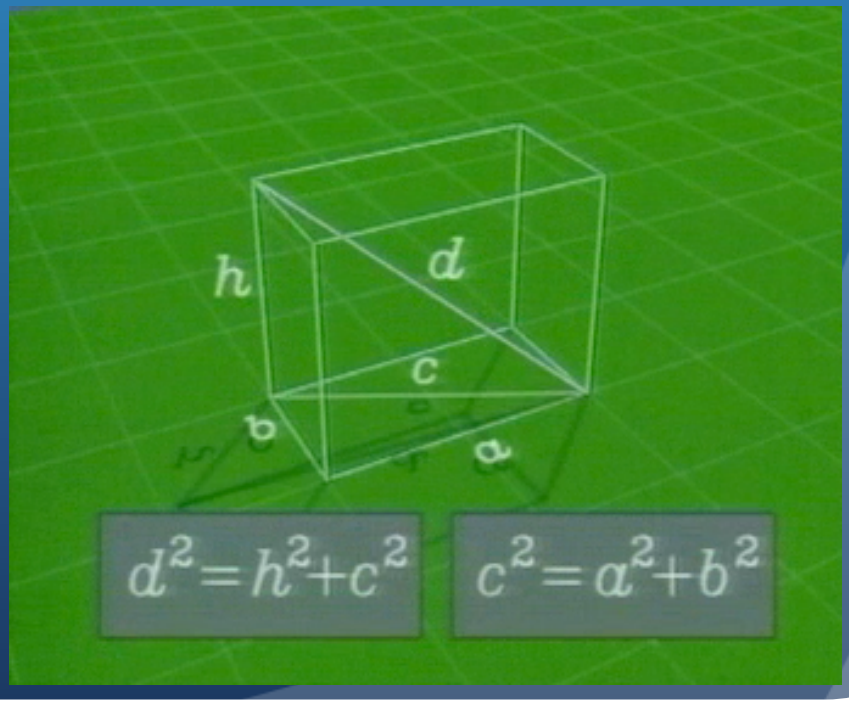
La **droite de régression** est la droite qu'on peut tracer dans le nuage de points qui représente le mieux la distribution à deux caractères étudiée. Il existe plusieurs manières de trouver l'équation de cette droite de régression. Outre l'utilisation des calculatrices graphiques et de certains logiciels, on peut calculer manuellement l'équation de la droite de régression.

Approche mathématique intuitive



Mesure de distance points-droite utilisée

- Longueur verticale
  - Incertitude face à la seconde composante des coordonnées
    Exemple : pluviométrie en fonction de l'altitude
    (var dépendante) (var indépendante)
- Au carré
  - Analogie avec le théorème de Pythagore 2D et 3D



Ce que nous avons appris dans cet atelier nous aide-t-il à mieux répondre à ceci ?

Les points, dans le graphique cartésien ci-dessous, représentent les coordonnées de nouvelles maisons dans un nouveau développement immobilier. L'entrepreneur du développement veut faire passer un réseau de fibres optiques sous terre le plus près possible de toutes ces maisons. Trouve l'équation linéaire qui représentera la position de la fibre optique souterraine que devrait construire l'entrepreneur de ce nouveau développement.

Maison	X	Y
A	10	30
B	25	20
C	50	70
D	65	60
E	120	90
F	40	45
G	80	90
H	100	70

http://www.alloprof.qc.ca/BV/Pages/m1378.aspx