http://irem.univ-reunion.fr/spip.php?article591

Découverte expérimentale de la droite de régression avec GeoGebra

- Lycée et post-bac
- Probabilités et statistiques
-

Date de mise en ligne : mercredi 29 août 2012

Il faut environ une heure à des étudiants de BTS pour redécouvrir expérimentalement

- **que la droite de régression obtenue par la méthode des moindres carrés passe par le point moyen du nuage ;**
- **que son coefficient directeur est le quotient de la covariance par la variance des abscisses ;**
- **pourquoi elle s'appelle Â« droite des moindres carrés Â» ;**
- **et ce qu'est le coefficient de corrélation.**

Une constatation préliminaire : environ la moitié de ces bacheliers (essentiellement STL) n'ont jamais vu GeoGebra de leur vie [1]...

Vidéoprojeter l'ordinateur d'un des étudiants aide grandement les autres, et les oblige à rester concentrés pendant toute l'heure.

Nuage de points

La création d'un nuage de points est rapide grâce au nommage automatique de GeoGebra. Cela met les étudiants en confiance (ils voient que GeoGebra est facile à utiliser et ne se découragent pas d'emblée) et permet de gagner du temps pour la suite, qui est plus longue. Pour faire le tout en une heure, il vaut mieux prendre 4 points ; mais par pure ambition, 5 points A, B, C, D et E ont été construits.

Pour construire la droite de régression obtenue par la méthode des moindres carrés, il suffit de sélectionner l'outil idoine (menu des droites construites), puis d'encercler les points avec un rectangle de sélection tracé à la souris :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH309/linreg1-b3ec0.png]

Après ça, il faut quand même faire un clic droit sur la droite [2] pour mettre son équation sous la forme y=ax+b, et en profiter pour la colorier en rouge. En examinant les propriétés de la droite, on découvre la syntaxe des listes :

```
{A,B,C,D,E}
```

qui servira par la suite.

Point moyen

Encore un raccourci utile : la possibilité d'additionner des points pour avoir directement le point moyen :

[http://irem.univ-reunion.fr/local/cache-vignettes/L198xH34/linreg2-bcfcf.png]

En mettant le point moyen G en vert (et en lui donnant une autre forme) pour le distinguer des points du nuage, on peut maintenant découvrir expérimentalement la propriété suivante :

> La droite de régression obtenue par la méthode des moindres carrés passe toujours par le point moyen du nuage.

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH248/linreg3-77dcd.png]

Moindres carrés

Ensuite, pour montrer d'où vient ce nom bizarre [3], on construit une autre droite (FH) en pointillés verts ci-dessous :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH257/linreg4-8821b.png]

Les points A, B, C, D et E ont été projetés sur la droite (FH) parallèlement à l'axe des ordonnées avec l'algorithme suivant :

1. On a tracé les droites verticales passant par A, B, C, D et E avec l'outil Â« parallèle Â» (en cliquant sur un des points puis l'axe des ordonnées) ; c'est là que les plus lents ou les plus maladroits commencent à ramer un peu ;
2. puis on a construit les intersections de ces droites verticales avec la droite (FH) ;
3. puis on a caché les droites verticales (en cliquant sur le bouton à gauche de leur nom dans la fenêtre algèbre) ;
4. puis on a construit les segments en mauve, que GeoGebra a nommés h, i, j, k et l ;
5. enfin on a caché les points d'intersection.

Là, on voit de tout !

La droite passant par B n'est pas toujours parallèle à l'axe des abscisses :

- Elle est parfois parallèle à la droite de régression.
- Elle est parfois parallèle à la droite (AB).
- Elle est parfois tracée au jugé, en construisant le point d'intersection avant la parallèle...
- Elle est parfois perpendiculaire à la droite (FH) [4].

Par ailleurs, certains étudiants préfèrent les segments aux droites et j'ai moi-même deux fois de suite construit les

intersections avec la droite de régression plutôt qu'avec la droite (FH) ; la fenêtre des propriétés de GeoGebra est d'une aide précieuse pour redéfinir après coup un objet et ceux qui en dépendent, et donc pour tous ceux qui, comme leur prof, font de grosses bêtises !

Pour éviter de surcharger la figure (et par manque de temps), on n'a pas tracé les carrés assis sur ces segments mauves, dont on cherche à minimiser la somme des aires. On s'est contenté de calculer algébriquement cette somme avec

```
sdc=h*h+i*i+j*j+k*k+l*l
```

L'abondance d'objets déjà construits rend malaisée la lecture de la fenêtre algèbre, et l'affichage de la somme des carrés a été introduit dans la figure sous la forme d'un texte dynamique, où la variable sdc est concaténée au texte par un Â« + Â» :

[http://irem.univ-reunion.fr/local/cache-vignettes/L368xH398/linreg5-c421b.png]

C'est cette étape qui a le plus épaté les étudiants. L'effet est assez puissant, il faut le reconnaître :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH266/linreg6-4dc8e.png]

Environ les deux tiers des étudiants ont réussi à minimiser la somme des carrés, et donc à découvrir que pour atteindre ce minimum, il fallait placer F et H sur la droite rouge. D'où son nom :

> La droite des moindres carrés est celle qui minimise la somme des carrés des distances verticales (c'est-à-dire les longueurs des segments verticaux) des points à la droite.

Covariance

Pour la suite (qu'un seul groupe a eu le temps de faire en entier), on peut rendre invisible tout ce qui a été construit à l'onglet précédent. La covariance a été définie dans le cours.

Comment ?

En utilisant la notation avec une barre pour la moyenne, on peut redéfinir les variances de x et y sans avoir à utiliser la notation Î£ :

[http://irem.univ-reunion.fr/local/cache-vignettes/L301xH198/covariance1-39d3a.png]

La définition de la covariance est alors naturelle ; son interprétation l'est beaucoup moins.

La trousse d'outils statistiques de GeoGebra est assez étendue pour inclure la covariance, et en plus, avec le raccourci de regrouper les x et les y comme coordonnées de points :

[http://irem.univ-reunion.fr/local/cache-vignettes/L260xH32/linreg7-d394f.png]

Sans trop chercher à comprendre ce qu'est exactement la covariance [5], on peut faire

```
c1=Covariance[{A,B,C,D,E}]
c2=Variance[{x(A),x(B),x(C),x(D),x(E)}]
c3=c1/c2
```

puis colorier c3 en rouge et le comparer avec le coefficient directeur de la droite de régression pour découvrir expérimentalement la propriété suivante :

> Le coefficient directeur de la droite de régression obtenue par la méthode des moindres carrés est le quotient de la covariance des abscisses et des ordonnées, par la variance des abscisses.

Ce qui fournit un algorithme permettant de calculer l'équation de la droite [6]

Corrélation

Dans le cours, le coefficient de corrélation a été défini comme le quotient de de la covariance par le produit des écarts-type :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH25/linreg8-2a499.png]

Mais GeoGebra a aussi un coefficient de corrélation, et peut même calculer directement celui des abscisses et ordonnées d'un nuage de points :

[http://irem.univ-reunion.fr/local/cache-vignettes/L284xH28/linreg9-11684.png]

La fin de l'activité a donc consisté à comparer ces deux nombres, et à évaluer l'impact de l'alignement des points sur la valeur du coefficient de corrélation :

Points très mal alignés :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH275/linreg10-619d0.png]

Points plutôt mal alignés :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH263/linreg11-178ff.png]

Points bien alignés :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH275/linreg12-74025.png]

Points bien alignés mais... :

[http://irem.univ-reunion.fr/local/cache-vignettes/L400xH254/linreg13-3ec0a.png]

Hacking

Tout ceci amène assez naturellement la question de savoir comment GeoGebra calcule les coefficients en question, et notamment si c'est avec les formules du cours. Or, GeoGebra étant un logiciel libre, il est tout-à-fait possible de répondre à cette question, tout simplement en consultant son code source [7]

Voici donc la manière dont GeoGebra calcule l'équation de la droite de régression (on constate le calcul simultané des moyennes et de la variance des x et de la covariance, dont les numérateurs sont nommés respectivement *sigmax*, *sigmay*, *sigmaxy* et *sigmaxx*) :

```
public final void compute() {
 double sigmax = 0;
 double sigmay = 0;
 double sigmaxx = 0;
 // double sigmayy=0; not needed
 double sigmaxy = 0;
 for (int i = 0; i < size; i++) {
 GeoElement geo = geoList.get(i);
 if (geo.isGeoPoint()) {
 double xy[] = new double[2];
 ((GeoPoint) geo).getInhomCoords(xy);
 double x = xy[0];
 double y = xy[1];
 sigmax += x;
 sigmay += y;
 sigmaxx += x * x;
 sigmaxy += x * y;
 // sigmayy+=y*y; not needed
 } else {
```

```
 g.setUndefined();
 return;
 }
 }
 // y on x regression line
 // (y - sigmay / n) = (Sxx / Sxy)*(x - sigmax / n)
 // rearranged to eliminate all divisions
 g.x = size * sigmax * sigmay - size * size * sigmaxy;
 g.y = size * size * sigmaxx - size * sigmax * sigmax;
 g.z = size * sigmax * sigmaxy - size * sigmaxx * sigmay; // (g.x)x +
 // (g.y)y +
 // g.z = 0
}<div class='code_download'   style='text-align: right;'> <a
href='http://irem.univ-reunion.fr/local/cache-code/0b2b9aebe5ed84be5adccb59c96a5df6.txt'   style='font-family:
verdana, arial, sans; font-weight: bold; font-style: normal;'>Télécharger
```

Si on fait abstraction du fait que l'équation de la droite est donnée sous forme homogène [8], l'équation de la droite de régression est bien établie à partir des deux informations suivantes :

1. Son coefficient directeur est le quotient de la covariance par la variance des x (plus précisément, le quotient de leurs numérateurs ; en effet ils ont le même dénominateur) ;
2. Elle passe par le point moyen.

Quant au calcul du coefficient de corrélation, lui aussi est fait d'après la définition du cours :

```
case STATS_PMCC:
 result.setValue((sumxy*sizex-sumx*sumy)/Math.sqrt((sumxx*sizex-sumx*sumx)*(sumyy*sizex-sumy*sumy)));<div
class='code_download'   style='text-align: right;'> <a
href='http://irem.univ-reunion.fr/local/cache-code/25cd9ee198e6cd4610ef3aaab3e7135c.txt'   style='font-family:
verdana, arial, sans; font-weight: bold; font-style: normal;'>Télécharger
```

Michael Borcherds économise le calcul d'une racine carrée en utilisant le fait que le produit des écarts-type est la racine du produit des variances. Et il économise les divisions par le nombre de points du nuage en gardant les numérateurs au lieu des moyennes.

---

[1] et souvent, pas d'autres logiciels de géométrie dynamique non plus.

[2] suivi de près par un clic gauche sur la gauche...

[3] c'est important que le nom de cet objet ait un sens, parce que d'autres droites de régression ont été vues au lycée, dont celle obtenue par l'algorithme de Mayer ; donner un sens à l'expression « moindres carrés » aide à comprendre le caractère unique de cette droite et à faire le lien avec la calculatrice.

[4] très bonne idée au demeurant : l'expression « moindres carrés » peut se comprendre de plusieurs manières, et la notion de distance d'un point à une droite est, fort heureusement, porteuse de sens...

[5] *a priori*, il suffit de savoir la calculer ; *a posteriori*, ce n'est même pas une exigence du programme, la seule chose qui sera demandée au BTS étant de savoir utiliser la calculatrice.

[6] la connaissance du coefficient directeur et d'un point, en l'occurrence le point moyen, suffit - théoriquement - à permettre de trouver l'ordonnée à l'origine.

[7] avec l'aide précieuse de Mathieu Blossier, de l'IREM de Rouen.

[8] depuis sa création en 2001, GeoGebra est un logiciel de Géométrie projective.