LE COURRIER DES LECTEURS

Les finesses de la régression

N *Dans le cours de statistique que G. Paturel nous propose dans le dernier numéro des Cahiers Clairaut, sur la régression linéaire, je lis dans l'introduction : "... le but est de trouver une relation linéaire entre deux grandeurs sur la forme Y= AX+B ..." puis, plus loin :"... il est donc nécessaire de ne tester que des relations linéaires (ce qui parait évident, compte tenu du nom de la méthode...) ..."*

Or, par la méthode des moindres carrés, exposée dans le paragraphe suivant, on peut ajuster non seulement des fonctions affines (représentées par des droites) sur un nuage de points , mais aussi des fonctions polynomiales :

*$Y = a_0 + a_1X + a_2 X^2 + a_3 X^3 + ...$ Les coefficients [...] se déterminent par un système linéaire de n équations à n inconnues. C'est la **linéarité** des paramètres ajustés [...] qui donne son nom à la régression. On peut même choisir*

$Y = a_0 + a_1 f_1(X) + a_2 f_2(X) +$ etc.. sous réserve que les fonctions f_1, f_2, etc. satisfassent à certaines conditions d'orthogonalité (et ne dépendent pas des a_i). Par exemple, on peut ajuster par une régression linéaire [..]

$Y = A.\cos BX$. La méthode est puissante et aurait pu être utilisée dans le traitement de la courbe de la figure 7 de

l'article de Philippe Jeanjacquot, sur la détermination du midi solaire, avec sans doute un gain substantiel de précision (ajustement par une parabole) ; ça doit être accessible à des Terminales [...] dans cet article, ne faudrait-il pas donner la longitude du lieu d'observation pour rendre pertinent cet écart de 14 ou 15 min avec le midi solaire, car l'azimut de 2° ouest correspond à 8 min de décalage horaire, l'équation du temps se rajoutant à ce décalage.

Daniel PASCAL

Merci pour ces commentaires judicieux. Effectivement, je me suis limité à la forme y=ax+b par simplicité. Mais il est évidemment possible de prendre des formes linéaires plus complexes (par exemple polynomiales : $y=ax + bx^2 +cx^3$...). J'avais signalé que x et y pouvaient être des fonctions (exemple de la relation PL), mais vous avez raison de faire remarquer que même une fonction trigonométrique rentre dans ce cadre. Cela se fait souvent avec des polynômes orthogonaux. En ce qui concerne l'article de Ph. JeanJacquot votre remarque est très judicieuse. Je l'ai transmise à l'auteur.

GP

Je ne suis pas du tout un spécialiste de statistiques et en particulier de la méthode des moindres carrés. Cependant, je voulais quand même vous apporter deux informations à propos de ce que vous avez écrit page 3 du cahier Clairaut 123. Vous écrivez : " Pourquoi le carré ? Simplement pour éviter qu'il y ait une compensation arithmétique. " Ce n'est certainement pas la bonne raison. Si c'était le cas, on aurait pu prendre la valeur absolue de la différence : |Y - Yi |.
Je me suis posé cette question il y a environ 35 ans, lorsque j'ai commencé à utiliser la méthode des moindres carrés pour traiter mes observations d'étoiles variables. A cette époque je fréquentais plusieurs chercheurs en Mathématiques pures et j'ai posé la question à deux d'entre eux. Il sont tombés d'accord pour me répondre que c'était très probablement dû aux propriétés très riches [...] de l'espace vectoriel des fonctions de carré intégrables. Je leur fais confiance !! On peut d'ailleurs comparer la droite obtenue par les moindres carrés et celle que donnerait les "moindres valeurs absolues". La méthode des moindres carrés est beaucoup plus sensible à la présence de points manifestement situés à l'écart du nuage de points approximativement alignés. A juste titre, vous envisagez le cas où les Xi ne sont pas entachés d'erreurs, puis le cas où ce sont les Yi et enfin le cas où les Xi et les Yi sont tous entachés d'erreurs possibles. Là, vous proposez de prendre la bissectrice des deux droites obtenues précédemment. En fait, on tombe ainsi sur la méthode générale des moindres carres et l'on cherche la droite qui passe au plus près des points (Xi, Yi). On minimise alors la somme des distances géométriques des points à une droite. Là encore, on peut minimiser une somme dont les termes sont ((X-Xi) B2 + (Y-Yi) B2) ou bien la somme des racines de ce genre de termes. Je n'ai pas cherché à voir la différence entre la droite ainsi obtenue et la bissectrice dont vous parlez, mais le calcul et sa représentation graphique doivent être très simples à faire en Maple.

Michel DUMONT

Merci pour vos commentaires éclairés. Selon "Méthodes statistiques" (Morice et Chartier, 1954) la probabilité d'un écart e_i est $\exp(-e_i^2/2s^2)$, dans le cas où les erreurs suivent une loi de Gauss. La probabilité de l'ensemble est donc le produit des probabilités, c'est-à-dire : $\exp(-\Sigma e_i^2/2s^2)$. Cette probabilité est maximum quand Σe_i^2 est minimum, d'où le critère. Dans l'article, je voulais seulement faire comprendre qu'on ne pouvait pas minimiser l'écart algébrique à la droite. Toujours d'après ma même source, la minimisation des valeurs absolues des écarts conduiraient à des calculs plus difficiles. Quant au choix de la bissectrice des deux régressions extrêmes dans le cas où les deux variables sont entachées d'erreur, c'est une simple approximation. On ne peut minimiser les distances géométriques que si les axes sont exprimés dans une même unité (sinon le résultat dépend des unités). Ma conviction est qu'un résultat n'est valable que s'il ne dépend pas de la méthode d'analyse. Je termine par une réflexion amusante de notre Président d'honneur Jean-Claude Pecker qui illustre le quotidien du chercheur :
"*L'astrophysique se ramène à deux opérations essentielles. D'abord (1) faire passer une droite par N points distribués au hasard, N étant aussi grand que l'on voudra; puis (2) faire passer par 2 points donnés une courbe d'ordre N, N étant aussi grand que l'on voudra*".

Les perles des enseignants, des astronomes et des autres...

La nécessité d'un exemple bien choisi.

Un professeur avait l'habitude de répéter à l'envie : 1 mètre carré c'est l'aire d'un carré de un mètre par un mètre. Un kilomètre carré, c'est l'aire d'un carré de 1 km par 1 km. Un bon élève en conclut qu'un carré de 5 mètres carrés devait être un carré de 5 mètres par 5 mètres.

Indépendamment, une bonne élève se fit cataloguer d'élève impertinente en ayant ri quand le professeur (non matheux) déclara qu'un évier de 0,50 m par 0,50 m avait une surface de 0,50 mètres carrés.

L'épilogue authentique de ces deux histoires authentiques, est que, plus tard, les deux bons élèves se marièrent et eurent beaucoup d'enfants... GP