Sozial- und Präventivmedizin Médecine sociale et préventive 26, 288 – 290 (1981)

ROBETH: un logiciel pour les procédés statistiques robustes

MARAZZI Alfio
INSTITUT UNIVERSITAIRE DE MEDECINE SOCIALE ET PREVENTIVE,
STATISTIQUE ET INFORMATIQUE MEDICALES, César-Roux 29, 1005 LAUSANNE

Résumé

Pendant les 20 dernières années, la théorie des statistiques robustes est devenue une branche très importante de la statistique mathématique. Certains concepts fondamentaux de cette théorie sont exposés ici à l'aide d'exemples. Une librairie de subroutines pour l'application des méthodes robustes a été développée dans le cadre d'un projet du Fonds National Suisse. Nous décrivons cette librairie dont le développement fait l'objet d'une partie de l'activité de notre Département de Statistique.

Introduction : les "outliers" et les statistiques "robustes"

Avec des données de type quantitatif, nous avons l'habitude de calculer des statistiques "classiques" telles que la moyenne arithmétique, les corrélations, les régressions par la méthode des moindres carrés. Nous allons voir par quelques exemples que des "observations aberrantes" peuvent avoir des effets importants sur les résultats de ces calculs.

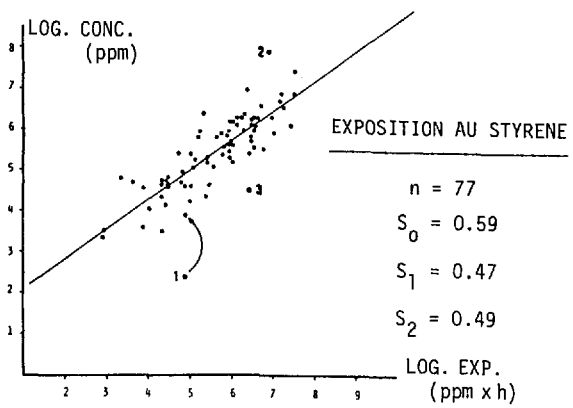


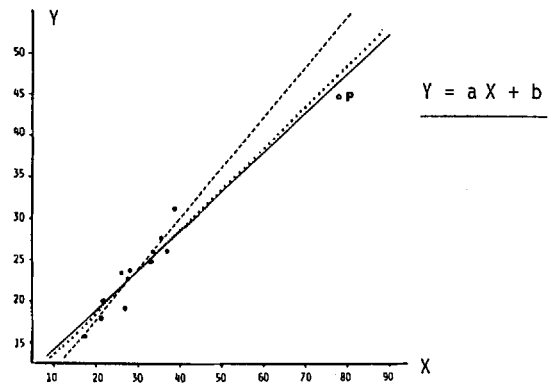
Figure 1

Le premier (voir figure 1) est un exemple réel de régression simple. Il s'agit d'une étude [1] sur l'exposition de travailleurs au styrène, solvant que l'on rencontre dans l'industrie du polyester. Chaque point représente un ouvrier. L'abscisse est le logarithme de l'exposition : c'est-à-dire du produit de la concentration moyenne du polluant à laquelle l'ouvrier a été soumis par la durée du travail. L'ordonnée est le logarithme de la concentration d'un métabolite du styrène (acide phenylglyoxylique) mesuré dans l'urine à la fin de l'exposition. Nous remarquons la présence de trois observations aberrantes numérotées par 1, 2 et 3. On les appelle en allemand "Ausreisser" et en anglais "outliers". Nous rencontrons ici les deux types d'outliers les plus fréquents dans nos études. Le no 1 est dû à une faute de transcription. Il a été possible de le corriger (comme indiqué sur la figure), en s'adressant simplement au responsable de l'étude. Les nos 2 et 3 sont à notre avis des points qui se situent un peu en dehors du groupe à cause de la présence de certains facteurs qui influencent les mesures, mais qui en général nous échappent (p. ex. : alimentation d'un sujet le jour avant l'examen, prise de médicaments, etc.). C'est le type d'outlier le plus fréquent parmi des données médicales. Dans cette étude particulière, le cas no 2 présentait également des valeurs exceptionnelles pour d'autres métabolites ainsi que pour la γGT. Le plus souvent, une fois qu'on connaît la raison de l'aberration, il paraît raisonnable de corriger ou d'enlever ces points du collectif. L'écart type S_0 indiqué sur la figure a été calculé avec le collectif complet. L'écart type S_1 a été calculé avec le collectif modifié : le point no 1 a été corrigé et les nos 2 et 3 éliminés. Il n'y a pas de différence entre les droites de régression, (pour cette raison seule la droite calculée sur le collectif complet est indiquée), mais on notera quand même une réduction de 25 % entre S_0 et S_1 . Cela pourrait avoir des conséquences décisives sur les conclusions d'un test statistique : avec une variance "gonflée", il serait par exemple plus difficile de rejeter l'hypothèse que l'ordonnée à l'origine est nulle.

Il n'est malheureusement pas toujours possible de procéder à une recherche très soigneuse des outliers. Les ordinateurs nous permettent de traiter de très grandes quantités de données et il est souvent impossible d'analyser chaque aberration pour des raisons de temps. Les moyens techniques pour la recherche des outliers peuvent aussi tomber en défaut. Pour les régressions, par exemple, il est coutume d'analyser les résidus des observations par rapport à la droite calculée. L'exemple suivant montrera le désavantage d'une telle procédure lorsqu'elle est basée sur des calculs classiques. En outre, il est bien connu que la recherche des outliers avec les moyens classiques et graphiques devient de plus en plus difficile, lorsque le nombre de paramètres augmente (régression multiple).

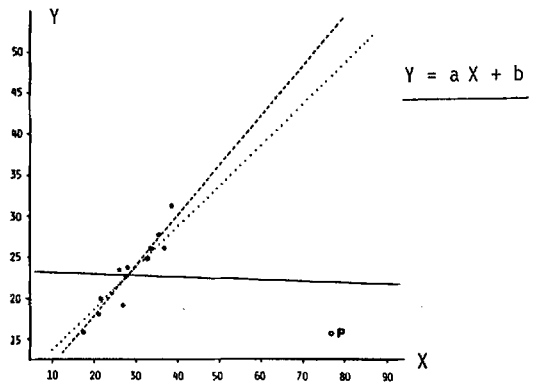
On a donc besoin de programmes qui puissent reconnaître automatiquement les outliers et les traiter convenablement. ROBETH répond à ce besoin. Il calcule des statistiques "robustes". Par cette expression, nous entendons (avec un sens un peu restreint par rapport à l'usage fait dans la littérature spécialisée), des statistiques qui ont les mêmes buts d'estimation, de description et d'analyse que les statistiques classiques que nous avons consi-

dérées, mais qui, contrairement à ces dernières, ne se laissent pas beaucoup influencer par les observations aberrantes. L'écart type S_2 de la figure 1 est le produit d'une estimation robuste.



$a_0 = .47$ $b_0 = 9.51$ $s_0 = 1.79$ $r_0 = .97$
$a_1 = .61$ $b_1 = 5.49$ $s_1 = 1.50$ $r_1 = .89$
$a_2 = .50$ $b_2 = 8.71$ $s_2 = 1.74$ $r_2 = .94$

Figure 2



$a_0 = -.01$ $b_0 = 23.16$ $s_0 = 4.90$ $r_0 = -.04$
$a_1 = .61$ $b_1 = 5.49$ $s_1 = 1.50$ $r_1 = .89$
$a_2 = .50$ $b_2 = 8.71$ $s_2 = 1.74$ $r_2 = .91$

Figure 3

Considérons maintenant un exemple artificiel (voir figure 2). L'abscisse du point P est aberrante, mais cependant ce point se trouve dans l'alignement des autres. Ce n'est donc pas un outlier pour ce problème et les résultats classiques sur le collectif complet (indice 0, droite continue), ceux qui ne tiennent pas compte du point P (indice 1, droite discontinue), et les résultats robustes (indice 2), droite pointillée) se ressemblent beaucoup (r_0 , r_1 et r_2 sont des estimations du coefficient de corrélation). Si nous déplaçons le point P dans le coin inférieur droit de la figure, il devient aberrant autant par son abscisse que par son ordonnée (voir figure 3). Tous les résultats des calculs classiques sont énormément influencés par un tel déplacement. On notera aussi que le résidu du point P par rapport à la droite continue est du même ordre de grandeur que les autres résidus. Cet inconvénient fait que, comme mentionné plus haut, certains outliers peuvent rester cachés. Par contre, les calculs robustes sur le collectif complet sont peu sensibles au déplacement de P . Evidemment, cela est vrai aussi pour les calculs qui rejettent le point P .

Quelques éléments de la théorie de l'estimation robuste

Dans les 20 dernières années, les statisticiens ont accordé une grande attention à l'estimation robuste. Des travaux pilotes ont été réalisés à l'Ecole Polytechnique de Zürich par les Professeurs Huber et Hampel (voir par ex. [2] et [3] et leurs listes bibliographiques). Ils ont proposé et étudié en particulier ce qu'on appelle aujourd'hui les estimateurs M , qui, parmi les estimateurs robustes, sont ceux qui ont eu le plus de succès. Cela grâce à leur efficacité, flexibilité et transparence théorique.

Nous considérons seulement l'estimation d'un paramètre de position. Nous supposons donc que les données y_i sont liées à un seul paramètre Θ par l'équation

$$y_i = \Theta + r_i \qquad (1)$$

où les r_i sont les erreurs de mesure. La théorie classique suppose que les erreurs sont indépendantes et identiquement distribuées selon la distribution de Gauss :

$$f(r) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2} . \qquad (2)$$

On peu estimer Θ par la méthode de vraisemblance maximum : l'estimateur $\hat{\Theta}$ est défini par la solution de l'équation

$$\sum f'(y_i - \hat{\theta})/f(y_i - \hat{\theta}) = 0 . \qquad (3)$$

On obtient la moyenne arithmérique : $\hat{\Theta} = \sum y_i/n$. Si une seule des observations tend vers l'infini, la moyenne arithmétique tend aussi vers l'infini. Elle n'est donc pas robuste.

La théorie robuste considère un modèle moins rigide. On fait l'hypothèse que seul un pourcentage $1 - \varepsilon$ des erreurs suit une distribution de Gauss et que le reste (les outliers), a une distribution inconnue h . La distribution des r_i est donc (modèle contaminé)

$$g(r) = (1 - \varepsilon)f(r) + \varepsilon h(r) . \qquad (4)$$

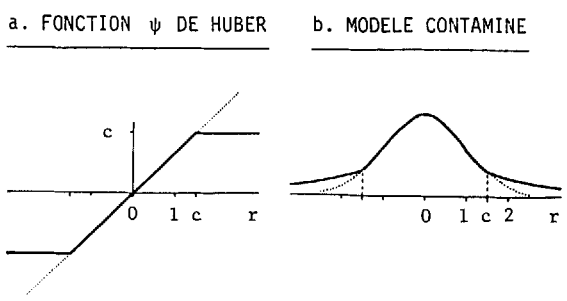
Pour estimer Θ on utilise un estimateur M , qui, par analogie avec l'estimateur de vraisemblance maximum est aussi défini par une équation :

$$\sum \psi(y_i - \tilde{\theta}) = 0 . \qquad (5)$$

Dans cette équation, la fonction ψ est tout d'abord arbitraire; il faut donc la choisir. Le choix se fait sur la base d'un problème d'optimalité que nous comparerons à un problème d'assurance. La moyenne arithmétique est optimale (c'est-à-dire au plus précis possible) pour le modèle idéal de Gauss. Elle n'est pas robuste et nous voulons nous assurer contre les accidents catastrophiques du type qu'on a vu (en d'autres termes, nous voulons que les effets des outliers ne dépassent pas une certaine limite). Il est clair que nous devons payer une prime. Nous admettons que la précision soit un peu diminuée dans le cas où le modèle idéal est vraiment approprié. Mais ce prix doit être le plus bas possible. Il y a

plusieurs façons de préciser et de formuler mathématiquement ce problème. Cela engendre des solutions ψ différentes, dont le prototype est la fonction de Huber (figure 4a). On peut démontrer que l'estimateur robuste correspondant est l'estimateur de vraisemblance maximum pour un modèle Gaussien dans sa partie centrale et exponentiel dans les queues (figure 4b). Tout cela se laisse généraliser à des situations plus complexes telles que la régression et le calcul de covariances.

La solution des équations du type (5) est réalisée à l'aide d'un programme d'ordinateur.



$$\psi(r) = \begin{cases} -c & \text{si } -c > r \\ r & \text{si } -c \leq r \leq c \\ c & \text{si } c < r \end{cases} \qquad g(r) = \begin{cases} (1-\varepsilon) f(c) e^{c(r+c)} & \text{si } -c > r \\ (1-\varepsilon) f(r) & \text{si } -c \leq r \leq c \\ (1-\varepsilon) f(c) e^{-c(r-c)} & \text{si } c < r \end{cases}$$

Figure 4

La librairie de sous-programmes ROBETH

Pour conclure, nous revenons maintenant à ROBETH. Son but est de mettre en circulation des programmes pour le calcul des statistiques robustes basées sur les estimateurs M . Il comprend actuellement à peu près 70 sous-programmes (subroutines) en FORTRAN. Dans leur réalisation, une attention particulière a été donnée à la portabilité et à la modularité. La documentation [4], [5], [6] et le mode d'emploi, ainsi que les programmes même peuvent être obtenus auprès de notre Département. Les méthodes implémentées sont pour le moment des estimateurs robustes pour :

- paramètres de position
- régressions simples et multiples
- matrices de covariances et de corrélations.

Nous implémentons actuellement des méthodes pour le choix des variables dans les problèmes de régression multiple.

BIBLIOGRAPHIE

[1] GUILLEMIN, M., BAUER, D., MARAZZI, A., MARTIN, B., (1981) Human Exposure to Styrene. IV. Industrial Hygiene Investigations and Biological Monitoring in the Polyester Industry. En préparation : sera soumis à Int. Arch. Occup. Env. Health.

[3] HAMPEL, F.R., (1973) Robust Estimation : A Condensed Partial Survey, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 27, pp. 87-104.

[2] HUBER, P.J., (1981) Robust Estimation, Wiley, New York.

[4] MARAZZI, A., (1980) ROBETH : A Subroutine Library for Robust Statistical Procedures, COMPSTAT 1980, Physika Verlag, Wien.

[5] MARAZZI, A., (1980) Robust Linear Regression Programs in ROBETH, ROBETH Document No. 2, Fachgruppe für Statistik, E.T.H., Zürich.

[6] MARAZZI. A., (1980) Robust Affine Invariant Covariances in ROBETH, ROBETH Document No. 3, Fachgruppe für Statistik, E.T.H., Zürich.

Zusammenfassung

ROBETH : eine Subroutinenbibliothek für robuste statistische Verfahren

In den letzten 20 Jahren wurde die Theorie robuster Schätzungen ein wichtiges Teilgebiet der mathematischen Statistik. Wir diskutieren hier einige Grundbegriffe dieser Theorie anhand einfacher Beispiele. Eine Subroutinenbibliothek zur Anwendung robuster Verfahren wurde im Rahmen eines Nationalfondsprojekts entwickelt. Wir beschreiben diese Bibliothek, an deren Weiterentwicklung in unserem Departement gearbeitet wird.

Summary

ROBETH : A Subroutine Library for Robust Statistical Procedures

In the past 20 years the theory of robust estimation has become an important topic of mathematical statistics. We discuss here some basic concepts of this theory with the help of simple examples. Futhermore we describe a subroutine library for the application of robust statistical procedures, which was developed with the support of the Swiss National Science Foundation.