Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Analyse critique

d'un ajustement linéaire

Rapport rédigé

par la

Société de Calcul Mathématique SA

Novembre 2018

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel 01 42 89 10 89. Fax 01 42 89 10 69. www.scmsa.eu
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

I. Présentation du besoin

Un Industriel veut évaluer les propriétés mécaniques d'une pièce métallique. Il dispose d'une centaine de prélèvements pris :

- A différentes étapes du process de fabrication ;
- En différents points de la pièce métallique.

La variable d'intérêt est une "résistance mécanique", notée R, dont on veut démontrer qu'elle ne descend jamais au-dessous d'un certain seuil.

Les variables mesurées sont au nombre de 7 : ce sont des propriétés chimiques (teneur en certains minerais), des propriétés physiques (températures, etc.).

L'Industriel fait l'hypothèse d'une dépendance linéaire entre les variables à expliquer et les variables explicatives (ou le logarithme de celles-ci), sous la forme d'une formule explicite :

$$R = a + bX_1 + cX_1^2 + dX_2 + eX_3 + fLog(X_4) + gLog(X_5) + hLog(X_6) + jLog(X_7) + reste \quad (1)$$

où R est la variable à expliquer et a, b, d, e, f, g, h, j sont des coefficients à déterminer.

Excel permet la détermination explicite des coefficients. Il suffit de mettre en colonne les différentes variables. Dans l'exemple numérique dont nous avons eu connaissance, l'ajustement apparaît comme extrêmement satisfaisant, puisque l'erreur relative moyenne est 1%.

L'erreur, pour chaque ligne, est la différence entre la valeur prédite pour R au moyen de l'ajustement et la valeur réelle de R. L'erreur relative est définie par la formule :

$$erreur\_relative(R) = \frac{|prédiction(R) - R|}{R} \times 100$$

et l'erreur relative moyenne est la moyenne de ces erreurs relatives, sur l'ensemble des mesures disponibles.

Au vu de ces résultats, l'Industriel pourrait vouloir négliger le reste, écrire :

$$R \approx a + bX_1 + cX_1^2 + dX_2 + eX_3 + fLog(X_4) + gLog(X_5) + hLog(X_6) + jLog(X_7) \quad (2)$$

et annoncer ceci :

Donnons à toutes les variables explicatives des valeurs arbitraires quelconques, non encore rencontrées, et calculons la variable de sortie au moyen de la formule (2), en utilisant les coefficients que nous venons de déterminer. Alors la variable de sortie sera évaluée correctement à 1% près.

La formule d'ajustement permet automatiquement de connaître une hiérarchie des variables explicatives : les variables les plus importantes sont celles qui ont le plus fort coefficient. Cette information est précieuse pour l'Industriel, mais la conclusion à partir de la formule (2) est incorrecte.

II. Critique méthodologique de cette approche

Même si, dans le cas particulier que nous voyons, l'ajustement donne un résultat très satisfaisant, la conclusion proposée plus haut n'est pas correcte.

1. Validation

Tout d'abord, l'ajustement n'a pas été validé. Il utilise, pour le calcul, la totalité des expériences disponibles. Un premier travail consiste en ceci : on retire par exemple un quart des expériences disponibles, on calcule l'ajustement sur les trois-quarts restants, et on se sert des coefficients ainsi calculés pour vérifier les valeurs prédites pour le quart mis de côté. Cette procédure, consistant à mettre de côté un certain nombre de valeurs expérimentales pour tester les résultats, s'appelle une "validation", et elle est essentielle en pareil cas.

Le nombre de valeurs que l'on peut mettre de côté pour la validation dépend évidemment du nombre de valeurs disponibles. Il faut pouvoir faire des moyennes sur les validations (il se peut très bien que le phénomène ait une forte variabilité), donc il ne faut pas se contenter de quelques valeurs retirées. Inversement, si on en retire trop, il en reste moins pour le calcul principal. Notre recommandation porte sur une fraction de 10 à 25 % pour la validation.

2. Extension du domaine de validité

Dans l'énoncé ci-dessus, nous disons "on peut donner aux variables explicatives des valeurs quelconques". Ceci n'est certainement pas correct. La formule d'ajustement linéaire définie en (2) n'est certainement pas valable pour des valeurs quelconques. Il y a un domaine de validité, qui dépend de la physique du problème.

Une analyse du domaine de validité de la formule (2) doit être fournie aux Autorités de Sûreté, mais, encore une fois, elle doit se déduire d'informations de nature physique. Tout ce que l'on peut faire, mathématiquement parlant, est de constater que l'ajustement est satisfaisant pour les données fournies ici. On ne peut savoir s'il en sera de même si certaines variables prennent des valeurs différentes.

3. Utilisation incorrecte de la formule

Reprenons la formule (2) en la simplifiant (on se limite à deux variables)

$$R = a + bX_1 + cX_1^2 + fLog(X_4) \tag{3}$$

Supposons que les calculs aient donné $f = 0.1$; l'Industriel en déduira que si X_4 est multiplié par 2, la valeur de R est modifiée de $0.1Log(2) \approx 0.069$, ce qu'il considérera comme négligeable. Autrement dit, pour lui, la variable X_4 ne joue pratiquement aucun rôle. Mais c'est parce qu'il a commis une erreur de logique, en introduisant un logarithme dans la formule (1). En effet, le logarithme est une fonction à croissance lente (qui convertit la multiplication en addition) et ce choix lui donnera l'impression fausse que X_4 n'intervient pas. Il aurait vraisemblablement eu une conclusion différente si, dans la formule (1), il avait fait intervenir X_4 et non $Log(X_4)$.

En d'autres termes, en introduisant des logarithmes dans la formule (1), l'Industriel a choisi de minimiser artificiellement l'importance de certaines variables, ce qui n'est certainement pas acceptable dans une démonstration de sûreté.

A l'inverse, le choix du terme cX_1^2 va accentuer la sensibilité du résultat par rapport aux variations de X_1.

III. Conclusions sur l'utilisation d'un ajustement linéaire

Une telle formule est très simple à mettre en œuvre (puisqu'il s'agit de fonctions préprogrammées dans Excel) et on peut toujours le réaliser, mais il ne peut servir de base à une démonstration de sûreté, remise aux Autorités, même si la formule a été validée sur un grand nombre de valeurs. En effet :

– Le domaine de validité de la formule est généralement inconnu ;

– La formule d'ajustement ne donne aucun intervalle de confiance sur le résultat.

Expliquons bien ce dernier point : si on utilise la méthode de manière prédictive, pour déterminer la valeur estimée de R à partir de valeurs connues des variables explicatives, on ne dispose pas d'un intervalle de confiance sur R, même si l'on dispose d'intervalles de confiance sur chacune des variables explicatives. Cela tient au fait que les coefficients de l'ajustement sont fournis de manière déterministe, sans ordre de grandeur quant à l'erreur commise.

L'utilisation de la formule d'ajustement ne donne aucune information sur la possible dispersion des variables sur lesquelles on travaille, aussi bien les variables explicatives que les variables de sortie : sont-elles concentrées ou dispersées ? Y a-t-il des valeurs aberrantes ?

En conclusion, pour répondre à la demande des Autorités de Sûreté, une approche probabiliste est nécessaire :

- Commencer par déterminer les lois de probabilité de toutes les variables (tracer les histogrammes) ;

- Déterminer les lois de probabilité conditionnelle (les dépendances) de la variable de sortie en fonction des divers paramètres ;

- En déduire une hiérarchisation de ces paramètres.

Voir notre fiche "méthodes robustes" pour plus de détails. Mais le but d'un dossier remis aux Autorités de Sûreté est, en définitive, l'estimation de certaines quantités avec pour chacune son intervalle de confiance. Les méthodes d'ajustement ne permettent pas l'évaluation de cet intervalle, à la différence de la méthode appelée "Hypersurface Probabiliste" (EPH), introduite par la SCM (voir le livre [PIT]).

IV. Références

Méthodes robustes : http://scmsa.eu/fiches/SCM_Methodes_robustes.pdf

[PIT] Olga Zeydina - Bernard Beauzamy : Probabilistic Information Transfer (en anglais), SCM SA, ISBN 978-2-9521458-6-2, ISSN 1767-1175, relié, 208 pages. Avril 2013.