Conférence spéciale

L'USAGE DES « MODÈLES LINÉAIRES GÉNÉRALISÉS » EN ANALYSE DE LA RÉGRESSION EST-IL ENCORE INDISPENSABLE AUX STATISTICIENS ?

Michel DELECROIX [(*)]*,Michel SIMIONI* [(**)]

[(*)] ENSAI
[(**)] INRA

L'analyse de la liaison entre variables observées est une des bases de l'analyse statistique. Elle présuppose la définition par le statisticien d'un modèle qui fixe la forme de ces liaisons, et requiert ensuite un travail classique d'ajustement aux données, via les techniques usuelles d'estimation et de tests. Le célébrissime « modèle linéaire gaussien » est l'exemple type de cette démarche.

L'utilisation de ce modèle admet de nombreuses limites. Une de ses extensions naturelles, permettant par exemple l'inclusion de variables expliquées catégorielles, est le « Modèle Linéaire Généralisé », au sens de Mc Cullagh et Nelder, dont relèvent par exemple les modèles Logit et Probit. L'utilisation de cet outil est devenue quasi-universelle, dans certaines branches de l'actuariat, par exemple, alors que certains types de données s'adaptent encore mal à ce modèle « généralisé », qui impose des choix *a priori* contraignants vis à vis de la loi qui engendre les observations, et dont il est difficile de tester l'adéquation aux données.

Pour échapper à cette difficulté inhérente à la recherche d'une bonne spécification du modèle, on peut essayer les techniques purement non paramétriques, qui n'imposent aucune condition sur la loi des observations. Il est cependant bien connu qu'elles sont mises en défaut dès que l'on considère plus de trois ou quatre variables explicatives, sauf à disposer de masses inhabituelles d'observations, ou bien lorsque les variables explicatives sont discrètes, par exemple.

Depuis une dizaine d'années, se sont par contre développées des techniques de modélisation qui, nécessitant au départ une seule des hypothèses de base du modèle linéaire généralisé (l'existence d'une « direction révélatrice »), disposent d'un champ d'utilisation beaucoup plus large que celui-ci. On se propose ici, après un bref rappel sur les hypothèses rendant pertinentes l'utilisation du « Modèle linéaire généralisé », de définir et discuter, en une première partie, l'hypothèse d'existence de « Directions révélatrices », et d'exposer les procédés essentiels d'estimation de la régression (« Dérivées Moyennées », M-estimation) qui en découlent.

Des études de cas seront ensuite présentées : on pourra comparer les résultats obtenus par les méthodes usuelles et les méthodes évoquées ci-dessus sur divers jeux de données, dont des fichiers INSEE concernant la participation des femmes au travail, ou l'analyse d'une fonction de gain…