Regression linéaire

Au chapitre précédent, nous avons utilisé divers outils graphiques (diagrammes en pointillés de Cleveland, diagrammes en boîte, histogrammes) pour explorer la forme de nos données (normalité), rechercher la présence de valeurs aberrantes et évaluer la nécessité de transformer les données.

Nous avons également abordé des méthodes plus complexes (coplot, scatterplot, pairplots, boxplots, histogrammes) qui ont permis de voir les relations entre une seule variable de réponse et plus d'une variable explicative. Il s'agit de la première étape essentielle de toute analyse qui permet au chercheur de se faire une idée des données avant de passer à des outils statistiques formels tels que la régression linéaire.

Tous les ensembles de données ne se prêtent pas à la régression linéaire. Pour les données de comptage ou les données de présences-absences, des méthodes générales de régression linéaire peuvent être utilisées.

Pour les données de comptage ou de présences-absences, la modélisation linéaire généralisée (GLM) est plus appropriée. Et lorsque les modèles paramétriques utilisés par la régression linéaire et la GLM donnent de mauvais résultats, techniques non paramétriques telles que la modélisation additive et la modélisation additive généralisée (GAM) sont susceptibles de donner de meilleurs résultats. Dans cet ouvrage, nous examinons une série d'outils adaptés à l'analyse des données univariées que l'on trouve couramment, y compris la régression linéaire, la régression linéaire partielle, le GLM, la modélisation additive, le GAM.

Les techniques telles que le GLM et le GAM sont plus difficiles à comprendre et à mettre en œuvre. C'est pourquoi nous commencerons donc par résumer brièvement les principes sous-jacents de la modélisation linéaire.

Regression linéaire bivariée

Le modèle de régression linéaire bivarié (c'est-à-dire à deux variables) est donné par la formule suivante:

$$Y_i = \alpha + X_i\beta + \epsilon_i$$

où est l'ordonnée à l'origine, est la pente et est le résidu, ou l'information qui n'est pas expliquée par le modèle. Ce modèle est basé sur l'ensemble de la population, mais comme expliqué ci-dessus, nous ne disposons que d'un échantillon de la population, et nous devons d'une manière ou d'une autre utiliser les données de cet échantillon pour estimer les valeurs de et de pour l'ensemble de la population. Pour ce faire, nous devons faire quatre hypothèses sur nos données qui permettront à une procédure mathématique de produire des valeurs estimées pour et . Ces estimateurs, appelés et , basés sur les données de l'échantillon agissent alors comme des estimateurs pour leurs paramètres de population équivalents, et respectivement. Les quatre hypothèses qui permettent d'utiliser les données de l'échantillon pour estimer les données de la population sont (i) la normalité, (ii) l'homogénéité, (iii) l'indépendance et (iv) la fixité de X.

Normalité

L'hypothèse de normalité signifie que si nous répétons l'échantillonnage plusieurs fois dans les mêmes conditions environnementales, les observations seront normalement distribuées pour chaque valeur de X.