Exercices supplémentaires – Régression linéaire

BLOC 1 en sciences informatiques

Exercice 1

Le fichier `ExRegrLin.txt` contient des données relatives à la taille et au poids de 20 individus. Malheureusement, un poids a été mal encodé.

1. Représenter le diagramme de dispersion du poids en fonction de la taille pour visualiser la donnée problématique. Déterminer le numéro de cette observation.

2. Ajouter au graphique précédent la droite de régression estimée par la technique des moindres carrés. Cette droite caractérise-t-elle bien la relation linéaire observée entre les deux variables ? Comment quantifier la qualité de l'ajustement ?

3. Les résultats précédents illustrent le manque de robustesse de la droite des moindres carrés et motive l'intérêt d'avoir recours à une alternative plus robuste.

   (a) Estimer à nouveau la droite de régression par la technique des moindres carrés en ne tenant pas compte de l'observation problématique. Pour ce faire, utiliser l'option `weights` de la fonction `lm()`, qui permet d'attribuer des poids $w_1, \ldots, w_n$ aux observations et d'ainsi minimiser

   $$\sum_{i=1}^{n} w_i r_i^2(a, b),$$

   où les $r_i^2(a, b)$ sont les résidus. Dans ce cas précis, puisqu'on veut ne pas tenir compte de l'observation détectée comme étant aberrante, on lui attribuera un poids nul et on attribuera un poids égal à 1 à toutes les autres observations. Ajouter la droite obtenue au graphique obtenu précédemment.

   (b) La régression LTS (*Least Trimmed Squares*) a été présentée lors du cours théorique. Pour rappel, cette technique consiste à "tronquer" la somme des carrés des résidus à minimiser en sélectionnant les h ($h < n$ à fixer) plus petits résidus (en valeurs absolues). Cette technique est disponible via la commande `ltsReg` de la librairie `robustbase`[1]. Il est possible de définir la *proportion de résidus à considérer* via l'option `alpha=...`, où `...` doit être remplacé par un nombre entre 0 et 1 (`alpha`$= h/n$). Appliquer cette régression LTS en tronquant un seul résidu et ajouter la droite de régression obtenue au graphique précédent. Commenter.

[1] Pour rappel, afin d'utiliser cette fonction provenant d'une librairie autre que celle de base, il faut préalablement charger le package dans `R` à l'aide de la commande `library(robustbase)`.

Exercice 2

Analyser les données de l'ensemble de données `stars.txt`. Ces données proviennent de l'astronomie et correspondent à des mesures effectuées sur 47 étoiles du cluster CYGOB1. Selon les scientifiques, un lien linéaire est attendu entre le logarithme de l'intensité de la lumière de l'étoile (variable dépendante) et le logarithme de la température à la surface de l'étoile (variable explicative).

1. Représenter le nuage de points et y ajouter la droite des moindres carrés. Commenter.

2. Représenter une boîte à moustaches des valeurs de la variable explicative et ajuster à nouveau la droite de régression par moindres carrés en attribuant un poids nul aux observations dont la valeur observée sur la variable explicative est extérieure à la boite. Ajouter cette droite au diagramme de dispersion et commenter.

3. Appliquer une régression LTS avec plusieurs valeurs de `alpha` de manière à visualiser l'impact du choix de ce paramètre sur la régression. Commenter.