TP nº 5 : Régression linéaire multiple 2

Exercice 1. On considère le jeu de données "profs". Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable Y),
- des résultats de 4 tests écrits donnés à chaque professeur (variables $X1$, $X2$, $X3$ et $X4$),
- du sexe (variable $X5$, avec $X5 = 0$ pour femme, $X5 = 1$ pour homme).

Le jeu de données est disponible ici :

`https://chesneau.users.lmno.cnrs.fr/profs.txt`

L'objectif est d'expliquer Y à partir de $X1$, $X2$, $X3$, $X4$ et $X5$. On considère le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \epsilon,$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres β_0, β_1, $\ldots$, β_5 et σ sont des réels inconnus.

1. Mettre les données sous la forme d'une data frame `w` en attachant les noms $X1$, $X2$, $X3$, $X4$, $X5$ et Y aux colonnes correspondantes.
2. Est-ce que les variables explicatives sont fortement corrélées entre elles ?
3. Donner des estimations ponctuelles des paramètres inconnus.
4. Donner la valeur prédite de Y lorsque $(X1, X2, X3, X4, X5) = (80, 150, 45, 44, 1)$.
5. Donner le R^2 et le R^2 ajusté. Que peut-on en dire ?
6. Est-ce que la régression est significative en $X2$?
7. Peut-on affirmer que $\beta_3 \neq -1.5$ au risque 5% ?
8. Donner un intervalle de confiance pour β_4 au niveau 95%.
9. On considère les hypothèses :

   $$H_0 : \beta_2 = \beta_3 = 0 \qquad \text{contre} \qquad H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0.$$

   Mettre en œuvre le test de Fisher :

   - en utilisant les formules du cours,
   - en utilisant des commandes adéquates.
10. On considère les hypothèses :

$$H_0 : \begin{cases} \beta_0 + \beta_1 + \beta_3 = 30 \\ \beta_2 + \beta_4 + \beta_5 = -30 \end{cases} \qquad \text{contre} \qquad H_1 : \beta_0 + \beta_1 + \beta_3 \neq 30 \text{ ou } \beta_2 + \beta_4 + \beta_5 \neq -30.$$

Mettre en œuvre le test de Fisher avec la fonction `linearHypothesis` de la librarie `car`.

11. Tracer l'ellipsoïde de confiance pour $(\beta_2, \beta_3)^t$ au niveau 95%.

12. Représenter le graphique des résidus. Identifier un point manifestement anormal avec la commande `identify`. Est-ce que les hypothèses standards semblent être satisfaites ?

13. Une nouvelle information nous parvient : en raison de problèmes de santé, les mesures de l'individu associé au point anormal ne sont pas fiables. Enlever cet individu du jeu de données et refaire l'étude précédente.

Exercice 2. On veut étudier la liaison entre la résistance à la traction du papier kraft (variable Y) en fonction du pourcentage de bois dur dans le lot de pâte à papier à partir de laquelle le papier a été produit (variable X). Le jeu de données est disponible ici :

`https://chesneau.users.lmno.cnrs.fr/kraft.txt`

1. Mettre les données sous la forme d'une data frame `w` en attachant les noms X et Y aux colonnes correspondantes.

2. Est-il judicieux d'envisager une liaison linéaire entre Y et X ?

3. On considère le modèle de *rlm* :
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$
avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres $\beta_0, \beta_1, \beta_2$ et σ sont des réels inconnus.
   (a) Donner des estimations ponctuelles des paramètres inconnus.
   (b) Représenter le nuage de points associé à (X, Y) et tracer la "courbe de régression" sur celui-ci.
   (c) Donner la valeur prédite de Y lorsque $X = 3$.
   (d) Donner le R^2 ajusté. Que peut-on en dire ?
   (e) Calculer le coefficient de corrélation (estimé) de X et X^2. Qu'en déduit-on ?

4. On considère le modèle de *rlm* :
$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \epsilon,$$
avec $X1 = X - \overline{x}$, $X2 = (X - \overline{x})^2$, $\overline{x}$ désigne la moyenne des observations de X et $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres $\beta_0, \beta_1, \beta_2$ et σ sont des réels inconnus.
   (a) Calculer le coefficient de corrélation (estimé) de $X1$ et $X2$. Qu'en déduit-on ?
   (b) Donner des estimations ponctuelles des paramètres inconnus.
   (c) Représenter le nuage de points associé à (X, Y) et tracer la "courbe de régression" sur celui-ci.
   (d) Donner la valeur prédite de Y lorsque $X = 3$.
   (e) Donner le R^2 ajusté. Que peut-on en dire ?
   (f) Représenter le graphique des résidus. Est-ce que les hypothèses standards semblent être satisfaites ?