M2 recherche TD4: Régression linéaire

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

On considère le modéle de régression

$$Y = 10 + x + 5\sin(x) + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, 4)$$

Pour des valeurs $x_i = i (i = 1, 2, \ldots, 30)$, générer un échantillon $y_1, \ldots, y_n$ ddi à partir du modèle précédent. Tracer les couples de points (x_i, y_i) ainsi que la vraie courbe. On oublie maintenant le modèle qui a généré ces données.

Exercice 1 (Modèle linéaire)
On suppose que le modèle est $Y = a + bx + \varepsilon$.

1. Quelles sont les formules permettant de trouver a et b ?
2. Que vaut la moyenne des résidus ? La moyenne des carrés ? Que vaut la moyenne des carrés des résidus si X et Y sont indépendants ? Si Y est totalement expliqué par X ?
3. Construire une fonction sous R calculant a et b.
4. L'appliquer aux valeurs x et Y. Tracer la droite $y = ax + b$ sur le graphique.
5. Utiliser la fonction `lm` de R pour retrouver le résultat.
6. Calculer la moyenne des carrés des résidus, vérifier qu'on retrouve bien la formule théorique.

La commande `lm`

- `reg<-lm(y~x)` : modèle $Y = aX + b$: estimation par moindres carrés.
- `lm(y~0+x)` : modèle passant par l'origine.
- `lm(y~1+x+I(x^2))` : modèle $Y = a_0 + a_1X + a_2X^2$: estimation par moindres carrés.
- `lm(y~x1+x2+x3)` : Régression multiple.
- `reg$call` : rappel du modèle de régression
- `reg$coefficients` : vecteur des coefficients estimés
- `reg$residuals` : résidus
- `reg$fitted.values` : estimation de Y.
- `summary(reg)` : test de non régression, etc
- `plot(reg)` : affichages divers

Exercice 2 ()
On suppose que le modèle est $Y = a + bx + c\sin(x) + dx^2 + \varepsilon$.

1. Utiliser R pour faire la régression.
2. Tracer la courbe obtenue. Commenter.
3. Quelle est la moyenne des carrés des résidus ?
4. Faire le test de nullité des coefficients.

   On change maintenant le modèle : on considère $Y = a + bx + c\sin(x) + \varepsilon$.
5. Utiliser R pour faire la régression. Tracer la courbe obtenue.
6. Quelle est la moyenne des carrés des résidus ?
7. Utiliser le critère AIC pour choisir le meilleur modèle.
8. Est-ce qu'on peut utiliser le critère R^2 ? Pourquoi ?