TP n° 9 : Introduction à la régression non-linéaire

Exercice 1. On veut étudier la liaison éventuelle entre la vitesse de formation d'une protéine (variable Y) et la concentration du substrat (variable X). En $n = 12$ expériences indépendantes, on observe les valeurs $(x_i, y_i)_{i\in\{1,\dots,n\}}$ de (X, Y) suivantes :

x_i	0.02	0.02	0.06	0.06	0.11	0.11	0.22	0.22	0.56	0.56	1.10	1.10
y_i	76	47	97	107	123	139	159	152	191	201	207	200

1. Représenter le nuage de points $(x_i, y_i)_{i\in\{1,\dots,n\}}$. À partir de celui-ci, expliquer pourquoi une liaison linéaire entre Y et X n'est pas adaptée.

2. Une solution est de transformer les variables. On considère le modèle de *rls* :

$$Y_* = \beta_0 + \beta_1 X_* + \epsilon,$$

avec $Y_* = 1/Y$, $X_* = 1/X$ et $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres β_0, β_1 et σ sont des réels inconnus.

   (a) Expliciter la liaison non-linéaire entre Y et X que l'on considère dans ce modèle.

   (b) Donner les *emco* ponctuels de β_0 et β_1.

   (c) Donner une estimation ponctuelle de σ.

   (d) Tracer la courbe de régression sur le nuage de points. Est-ce que cette courbe ajuste bien le nuage de points ?

3. Une autre solution est de considérer directement un lien non-linéaire entre Y et X. Dans ce contexte, les biologistes utilisent le lien de Michaëlis-Menten : $Y = f_{k,v}(X)$, avec

$$f_{k,v}(x) = \frac{vx}{k + x}.$$

La modèle de régression non-linéaire associé est : $Y = f_{k,v}(X) + \epsilon$, avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres v, k et σ sont des réels inconnus. Ainsi, pour tout $i \in \{1, \dots, n\}$ et $X = x_i$ fixé, y_i est une réalisation de $Y_i = f_{k,v}(x_i) + \epsilon_i$, où $\epsilon_1, \dots, \epsilon_n$ sont n *var iid* suivant chacune la loi normale $\mathcal{N}(0, \sigma^2)$.

L'estimation par les moindres carrés ordinaires consiste à trouver $\widehat{k}$ et $\widehat{v}$ tels que

$$(\widehat{k}, \widehat{v}) = \underset{(k,v)\in\mathbb{R}^2}{\text{Argmin}} \sum_{i=1}^{n} (Y_i - f_{k,v}(x_i))^2.$$

Cette minimisation se fait par le biais d'un algorithme qui requiert des valeurs initiales (k_0, v_0) pas trop éloignées de la solution supposée. On observe que $\lim_{x\to\infty} f_{k,v}(x) = v$ et $f_{k,v}(k) = v/2$, donc k est l'abscisse de la courbe quand on se trouve à l'ordonnée $v/2$. D'après le nuage de points on peut prendre $k_0 = 0.05$ et $v_0 = 200$.

Reproduire et comprendre l'enjeu des commandes suivantes :

```
x = c(0.02, 0.02, 0.06, 0.06, 0.11, 0.11, 0.22, 0.22, 0.56, 0.56, 1.1, 1.1)
y = c(76, 47, 97, 107, 123, 139, 159, 152, 191, 201, 207, 200)
w = data.frame(x, y)
plot(w)
reg = nls(y ~ v * x / (k + x) , data = w, start = list(v = 200, k = 0.05))
summary(reg)
library(ellipse)
plot(ellipse(reg, which = c(1, 2)))
beta = coef(reg)
plot(w)
curve(beta[1] * x / (beta[2] + x), add = T)
e = residuals(reg)
plot(e)
qqnorm(e)
```

4. Représenter les 2 courbes de régression obtenues aux questions 1 et 2 sur le même nuage de points. Quelle est celle qui ajuste le mieux le nuage de points ?

Exercice 2. La mesure du niveau d'anticorps d'un serum de vache peut d'effectuer ainsi : On mesure la densité optique de solutions (variable Y) correspondant à diverses dilutions du sérum (l'opposé du logarithme népérien de la dillution sera une variable X). En $n = 9$ expériences indépendantes, on observe les valeurs $(x_i, y_i)_{i\in\{1,\dots,n\}}$ de (X, Y) suivantes :

x_i	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y_i	1.81	1.80	1.75	1.60	1.20	0.60	0.30	0.10	0.08

L'expérimentateur souhaite ajuster le nuage de points par la courbe d'équation :

$$f_\theta(x) = \theta_2 + \frac{\theta_1 - \theta_2}{1 + e^{\theta_3 + \theta_4 x}},$$

avec $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, $\theta_1 > 0$, $\theta_2 > 0$, $\theta_1 > \theta_2$ et $\theta_4 > 0$. La modèle de régression non-linéaire associé est : $Y = f_\theta(X) + \epsilon$, avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres θ_1, θ_2, θ_3, θ_4 et σ sont des réels inconnus.

1. Créer une data frame `w` qui comprendra une colonne `x` avec les valeurs de x et une colonne `y` avec les valeurs de y.

2. Construire un objet R qui est le résultat d'une régression non-linéaire sur les données `w` *(on prendra pour valeurs initiales :* $\theta_0 = (1.82, 0.082, -8.975834, 2.761795)$*)*. Quelles sont les estimations des paramètres inconnus ?

3. Tracer le nuage de points et la courbe de régression obtenue.

4. Changer les valeurs initiales de l'algorithme pour le tester. Par exemple, on pourra prendre $\theta_0 = (30, -20, 0, 32)$.