INSA – GM 5 Année 2018-2019
ANTOUN Cyril
ROUFFIAC Jean-Eudes

**Régression Non Linéaire avec Applications sous R**

Rapport de TP n°4

Titre : *Régréssion non linéaire paramétrique*

# Introduction

L'objectif de ce TP est d'expérimenter quelques techniques statistiques dans le cadre des modèles de régression non linéaires paramétriques. Les données étudiées représentent 217 mesures du taux de croissance d'une bactérie à différentes températures. On propose de modéliser l'influence de la température T, sur le taux de croissance Y en phase exponentielle des micro-organismes à l'aide du modèle suivant :

$$Y_i = \mathrm{f}(t_i,\theta) + \epsilon_i$$

où :

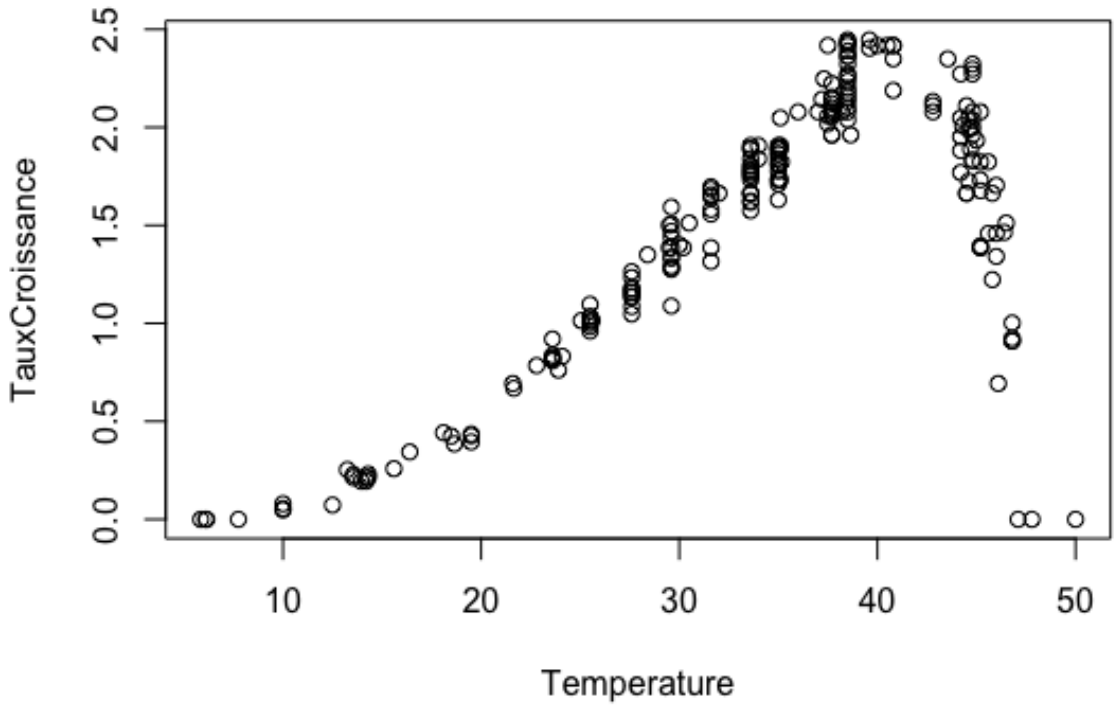
- $f(\mathrm{t}_i,\theta)$ décrit la relation entre la température et le taux de croissance
- les erreurs $(\epsilon_i)$ sont des variables aléatoires que l'on suppose indépendantes et de même loi $\mathcal{N}(0,\sigma^2)$

Pour ce problème, un choix possible de fonction f est :

$$f(t,\theta) = \begin{cases} 0 & \text{si } t \notin [T_{min}, T_{max}] \\ \frac{Y_{opt}(t-T_{max})(t-T_{min})^2}{(T_{opt}-T_{min})[(T_{opt}-T_{min})(t-T_{opt})-(T_{opt}-T_{max})(T_{opt}+T_{min}-2t)]} & sinon \end{cases}$$

avec $\theta = (T_{min}, T_{max}, T_{opt}, Y_{opt})$ où $T_{min}$ représente la température en deçà de laquelle il n'y a plus de croissance, $T_{max}$ la température au delà de laquelle il n'y a plus de croissance, $T_{opt}$ la température pour laquelle le taux de croissance atteint son maximum $Y_{opt}$. C'est le modèle dit des températures cardinales.

Dans un premier temps, on représente le taux de croissance en fonction de la température. On obtient le graphique suivant :



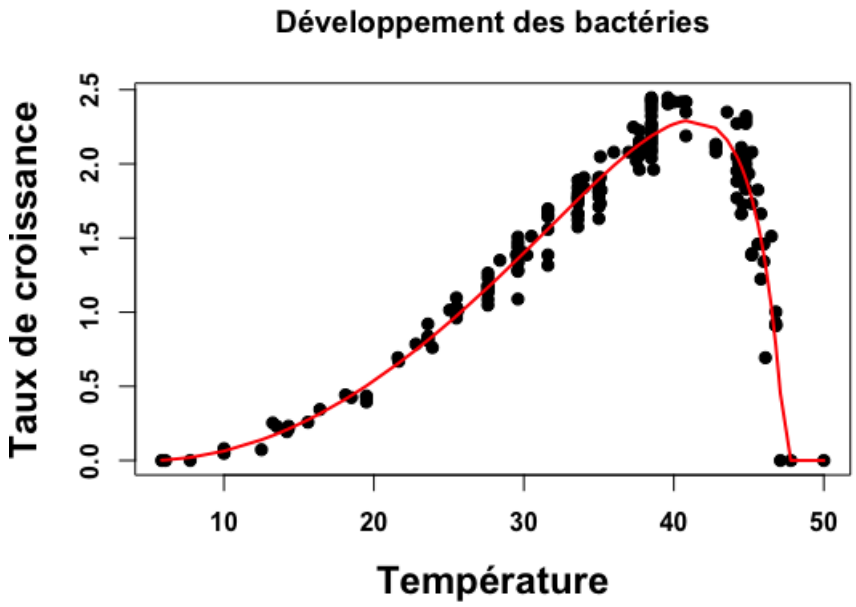

On remarque que le modèle n'est pas linéaire et on ne repère pas de modèle déjà étudié. Le nuage de point présente une première phase où le taux de croissance augmente relativement lentement pour une température comprise entre 2 et $40°C$. A cette température, le taux de croissance atteint son maximum d'une valeur d'environ 2.4. S'ensuit alors une deuxième phase où le taux de croissance chute nettement pour une température supérieure à $40°C$.

On veut maintenant estimer les paramètres du modèle envisagé. On se servira d'un échantillon d'apprentissage contenant 85% des données initiales. Grâce au nuage de points on donne les conditions initiales de l'algorithme suivantes : $T_{min} = 2$, $T_{max} = 50$, $T_{opt} = 40$ et $Y_{opt} = 2{,}4$. Avec le logiciel R on trouve les résultats suivants :

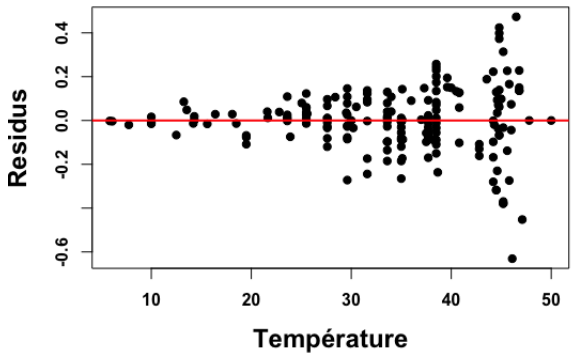
| | |
|---|---|
| Tmin | 4.88452 |
| Tmax | 47.43952 |
| Topt | 41.26763 |
| Yopt | 2.29416 |

On remarque que la convergence s'effectue en 5 itérations. Les faibles p-values calculées pour chaque coefficient nous permettent de déduire que les coefficients sont significativement différents de 0 au risque 5%. Enfin, on relève un faible Residual Standard Error de 0.15.

On représente alors la fonction f avec les paramètres estimés sur le nuage de points.

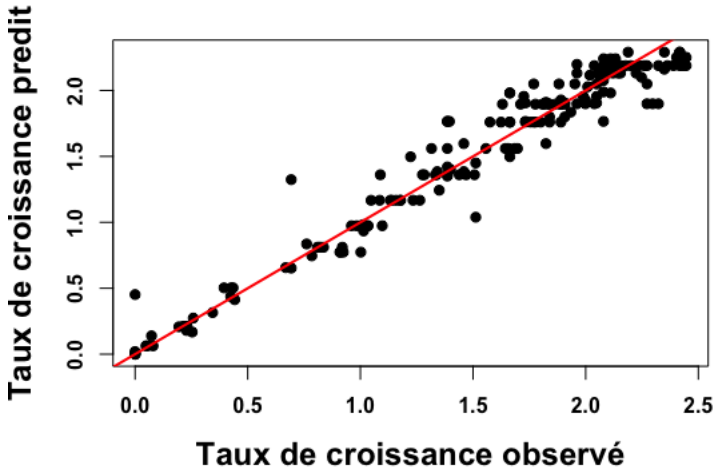



On remarque que la fonction estime très bien le nuage de points. Analysons maintenant la qualité de l'ajustement. On étudie donc dans un premier temps les résidus non normalisés.



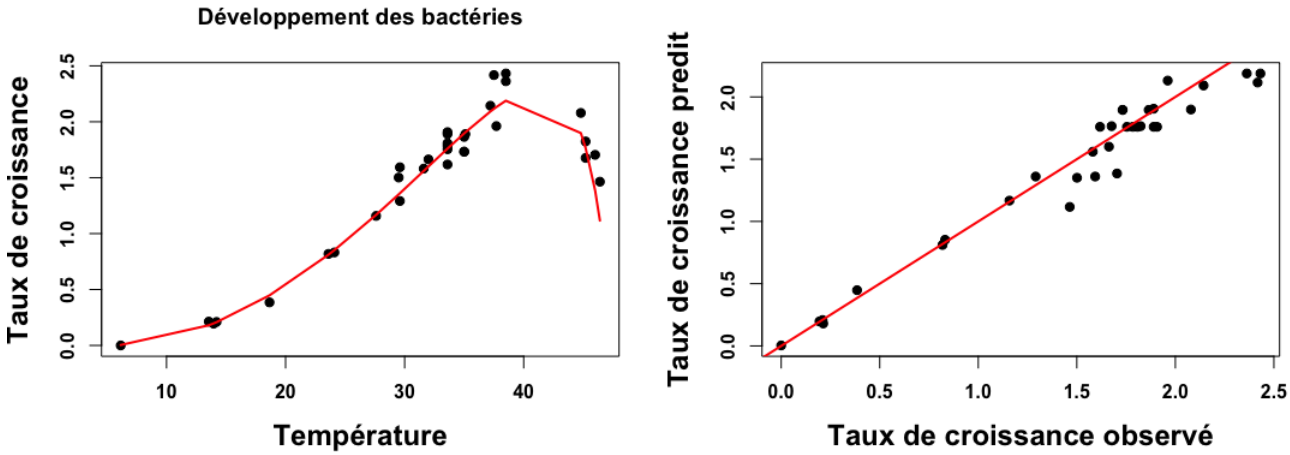
On remarque grâce à ce graphique que les résidus sont équitablement répartis autour de zéro. Ils sont très faibles pour une température inférieure à $25°C$ mais deviennent plus grands au delà de cette température. On voit clairement que l'hypothèse d'homoscédasticité n'est pas vérifiée.

Étudions maintenant les performances du modèle en examinant le nuage de points observé-estimé :




On remarque que le modèle ajuste bien les données. Pour des forts taux de croissance observés (supérieur à 2), le modèle sous estime légèrement les données alors que pour un taux de croissance compris entre 1 et 2 il sur-estime légèrement les données.

Étudions maintenant les performances du modèle sur l'échantillon test, qui contient 15% des données initiales.

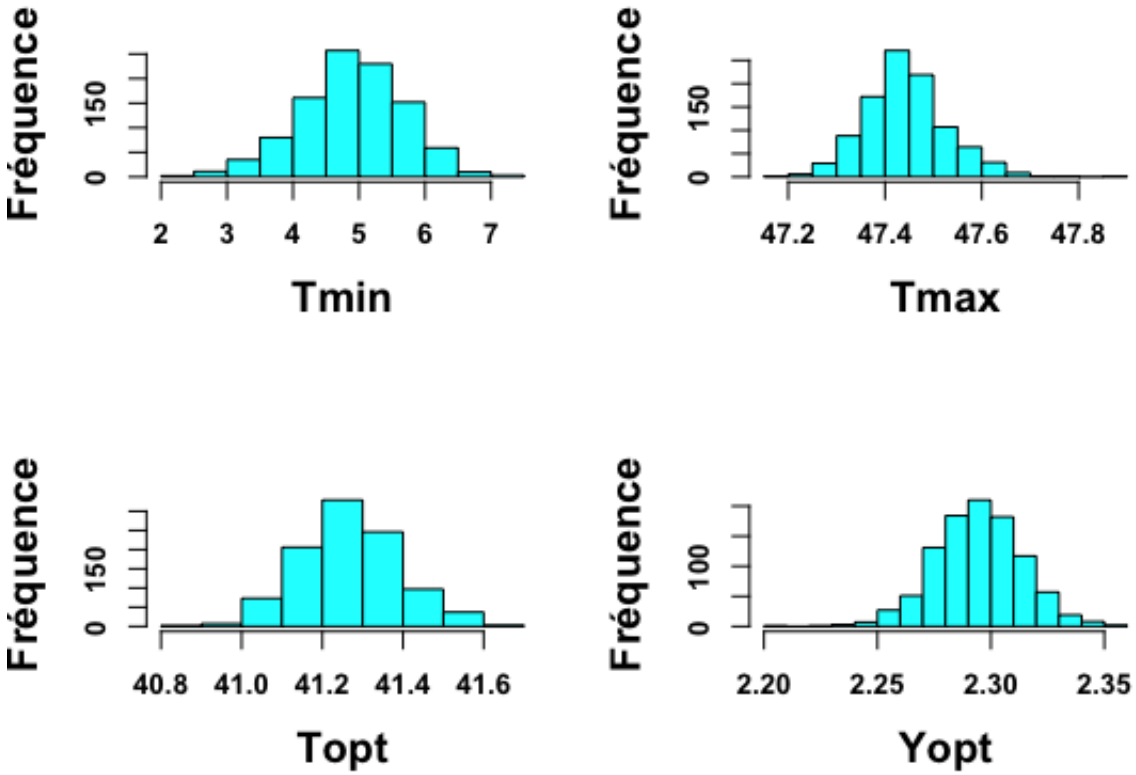



On remarque sur le premier graphique que le modèle estime très bien le nuage de points des données de l'échantillon test. Sur le deuxième graphique, on a représenté le taux de croissance prédit en fonction du taux de croissance observé. On voit que le modèle ajuste bien les données. De même que pour l'échantillon d'apprentissage, pour de forts taux de croissance observés (supérieur à 1.5), le modèle sous-estime légèrement les données. Enfin, la valeur du RMSE égale à 0.14 et celle du MAE à 0.10 montrent que le modèle ajuste donc bien les données de l'échantillon test.

On va maintenant utiliser la méthode de réechantionnage bootstrap sur les résidus pour calculer les erreurs standard pour les paramètres $\hat{\theta}_i$. On crée donc 1000 nouveaux échantillons et on calcule les erreurs standard de chaque paramètre. On obtient le tableau suivant dans lequel on a rajouté le biais des estimations :

| Paramètre | Biais | Erreur standard de R | Erreur standard Bootstrap |
|---|---|---|---|
| $T_{min}$ | 0.034 | 0.820 | 0.786 |
| $T_{max}$ | 0.002 | 0.085 | 0.083 |
| $T_{opt}$ | 0.004 | 0.127 | 0.123 |
| $Y_{opt}$ | 0.000 | 0.019 | 0.019 |

Toutes les erreurs standards sont faibles et semblables d'une méthode à l'autre. De plus, les ordres de valeur des biais sont faibles comparées aux valeurs obtenues des paramètres. On représente maintenant la distribution des estimations bootstrap obtenues des quatre paramètres. On obtient les histogrammes suivants :




Visuellement, les quatre distributions semblent gaussiennes. Cependant, après avoir effectué un test de Shapiro-Wilk, on rejette la normalité de la distribution des estimateurs de $T_{max}$ et $T_{opt}$ au risque 5% et on accepte la normalité de la distribution des estimateurs de $T_{min}$ et $Y_{opt}$.

On construit maintenant un intervalle de confiance à 95% avec la méthode des pourcentiles et celle de Student pour chacun des paramètres :

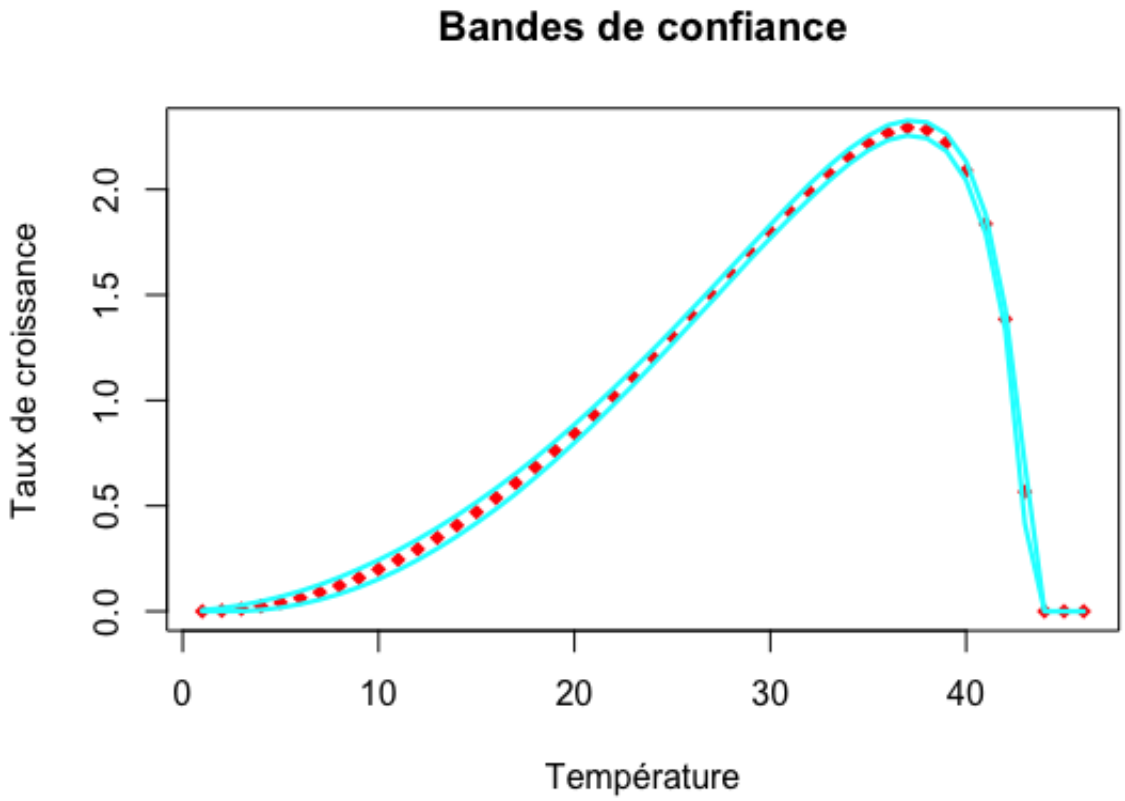
| Paramètre | Estimation | Intervalle avec pourcentiles | Intervalle avec Student |
|---|---|---|---|
| $T_{min}$ | 4.88 | [3.25 ;6.30] | [3.28 ;6.48] |
| $T_{max}$ | 47.44 | [47.29 ;47.61] | [47.27 ;47.61] |
| $T_{opt}$ | 41.27 | [41.03 ;41.52] | [41.02 ;41.52] |
| $Y_{opt}$ | 2.29 | [2.26 ;2.33] | [2.26 ;2.33] |

Les deux méthodes donnent des intervalles similaires pour tous les estimateurs. On pouvait s'y attendre pour $T_{min}$ et $Y_{opt}$ au vu du caractère gaussien de ces estimateurs mais cela est plus étonnant pour les deux autres paramètres qui ne sont pourtant pas gaussiens.

L'intervalle de confiance du paramètre $T_{min}$ apparait assez large. En revanche, pour les

paramètres $T_{max}$, $T_{opt}$ et $Y_{opt}$ les intervalles de confiance sont très ressérés autour de l'estimation. On en déduit donc que les estimations calculées sont fiables.

Pour finir, on construit une bande de confiance au niveau 95% pour le modèle obtenu :




On peut remarquer que la bande de confiance n'est pas très large. Elle s'élargit pour des températures autour de $15°C$ et $40°C$, qui sont les températures où les variations du taux de croissance sont les plus fortes.