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

Régression Non Linéaire avec Applications sous R

Rapport de TP n°5

Titre : *Estimation non-paramétrique de la fonction de régression*

1 Première partie : Etude par simulations

1.1 Loi normale

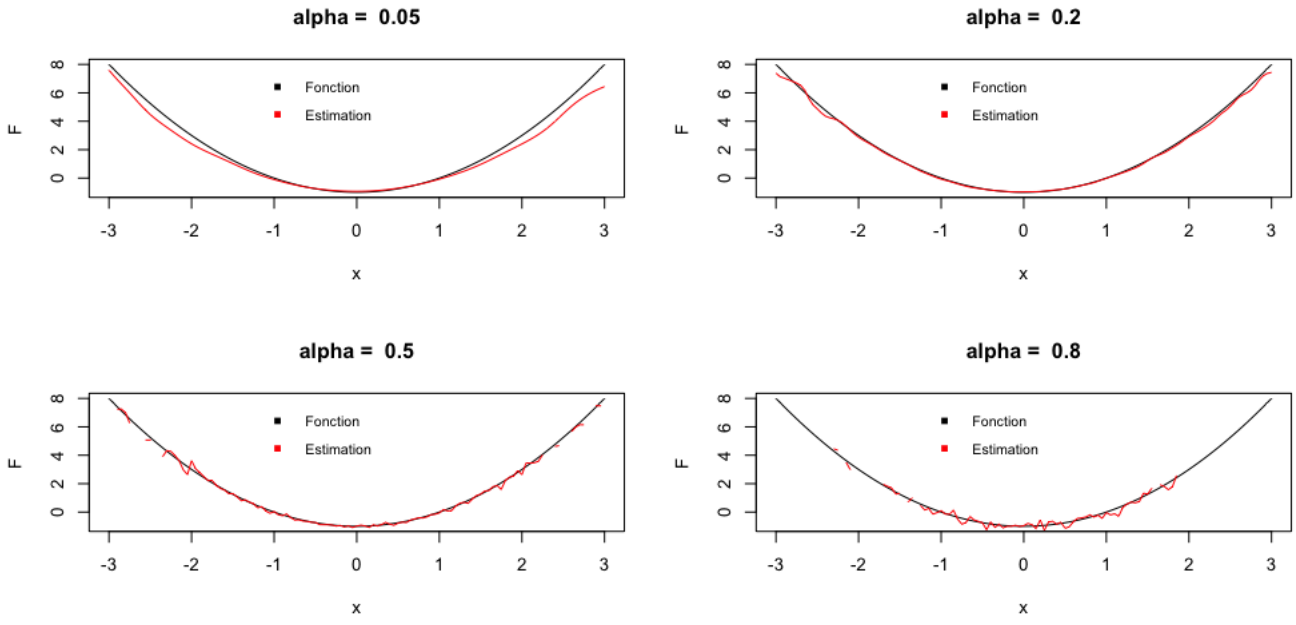
Dans cette première partie nous simulons 1000 réalisations$(x_i, y_i)_{1\leq i\leq 1000}$ du couple (X,Y) de variables aléatoires, sachant que X suit une loi normale centrée réduite et

$$Y = f(X) + \epsilon.$$

où :

- $f(X) = (x^2 - 1)$
- ϵ suit une loi normale N(0, 0.25)

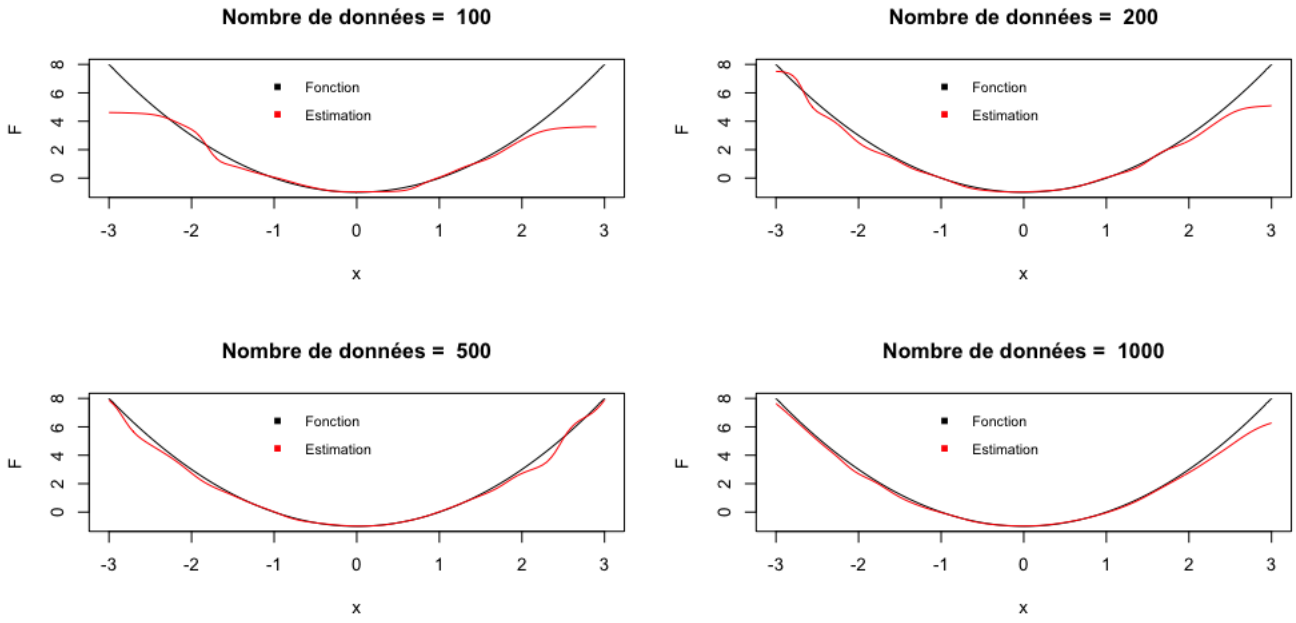
Dans un premier temps, on étudie l'influence d'une fenêtre de la forme $\widehat{\sigma_X} n^{-\alpha}$ où $\alpha \in [0; 1]$ et $\widehat{\sigma_X}$ est l'estimation de l'écart type de X. On fixe la taille n de l'échantillon à 1000. Nous avons testé plusieurs valeurs de α mais nous allons en présenter que quatre avec les graphiques suivants :



On remarque que pour un α petit, inférieur à 0.2, l'estimation est lisse mais on observe un biais trop important. Inversement, pour un α grand, supérieur à 0.2, le biais est faible mais la courbe n'est plus lisse. Une valeur optimale de α semble être proche de 0.2.

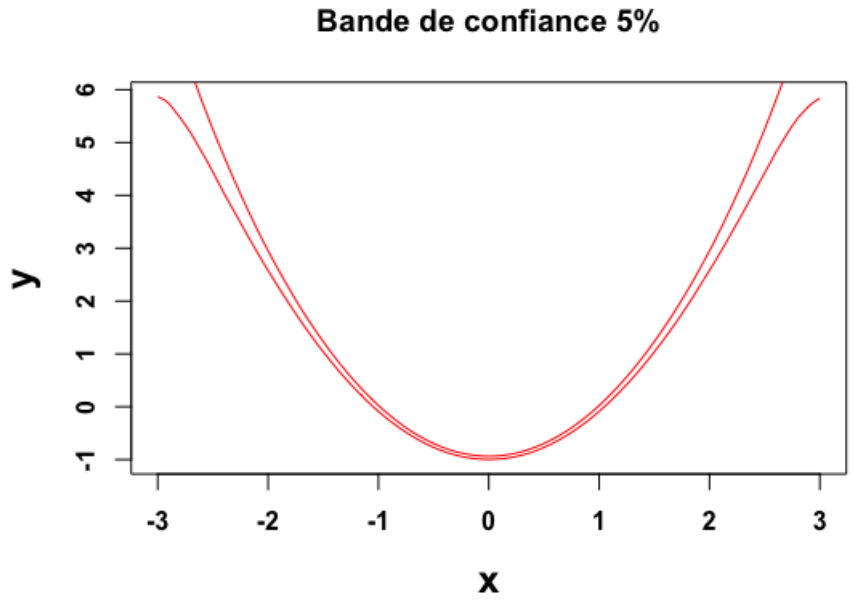
On cherche maintenant à étudier, en fonction de la taille n de l'échantillon, la qualité de l'estimation de la fonction f par l'estimateur de Nadaraya-Watson. On prend n=100,

200, 500 et 1000. On utilise un noyau gaussien et la fenêtre optimale. Les résultats sont les suivants :



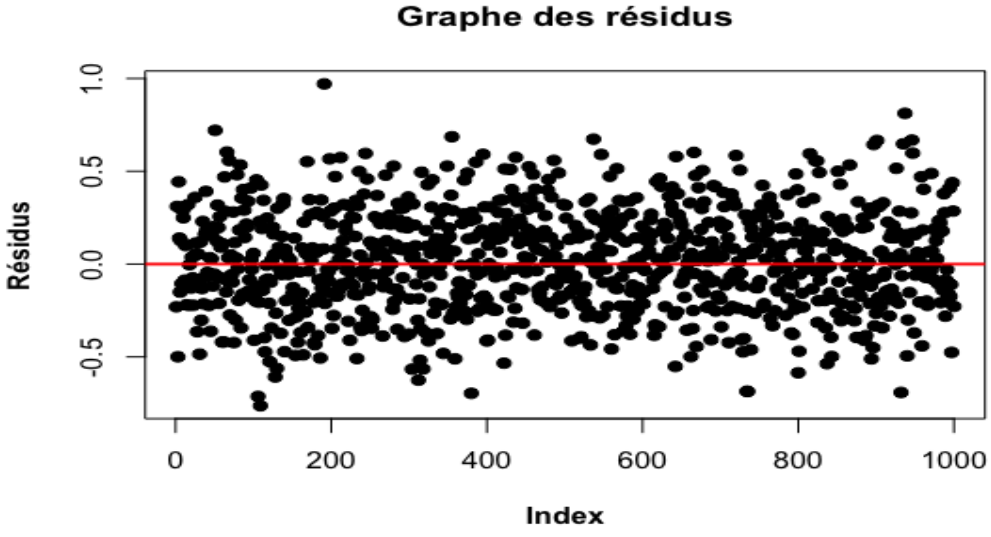
Nous remarquons que pour des faibles valeurs de n, inférieures à 200, l'estimation n'est pas satisfaisante. Elle approxime mal la fonction f, notamment pour des valeurs extrémales de x. Cependant, pour de grandes valeurs de n l'estimation approxime correctement la fonction f. De manière générale, pour x proche de zéro l'estimation apparait très précise mais l'est beaucoup moins lorque x se rapproche des bornes -3 et 3.

On va maintenant utiliser la méthode de Monte-Carlo pour obtenir une bande de confiance de l'esimation de f. Nous utilisons la méthode des percentiles simples et effectuons 2000 simulations pour déterminer en toute valeur de t de notre intervalle d'étude, l'intervalle de confiance pour f(t). On conserver nos valeurs de n et α trouvés aux questions précédents, à savoir 1000 et 0.2. On présente le résultat sur le graphique ci-dessous :



On remarque que la bande est très resserée pour x proche de 0. Elle est plus large lorsque x est compris dans les intervalles [-3,-2] et [2,3] mais reste tout de même resserée. On peut conclure que l'estimation de f est bonne.

En conclusion, afin d'apprécier la qualité du modèle estimé, nous faisons une étude rapide des résidus. Ainsi le graphe des résidus est donné ci-dessus :



Les résidus sont faibles et répartis équitablement autour de 0. L'hypothèse d'homocédasticité semble validée. De plus, le test de Shapiro-Wilk donne une p-value de 0,37 ce qui nous permet d'accepter au risque 5% le caractère gaussien des résidus. Enfin la moyenne quadratique des résidus est de 0.257, soit très proche de l'écart-type de ϵ qui est de 0.25. Ces éléments confirment que l'estimation de la fonction f par l'estimateur de Nadaraya-Watson est de très bonne qualité.

1.2 Loi uniforme

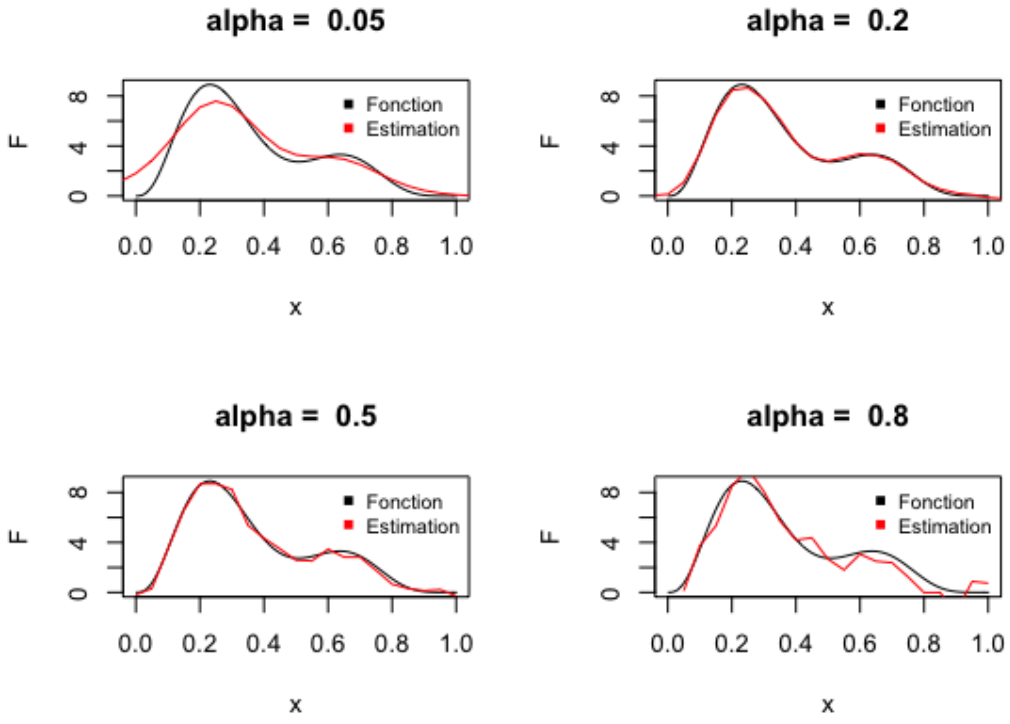
A nouveau nous simulons 1000 réalisations $(x_i, y_i)_{1 \leq i \leq 1000}$ du couple (X,Y) de variables aléatoires, sachant que cette fois X suit une loi uniforme [0,1] et

$$Y = f(X) + \epsilon.$$

où :

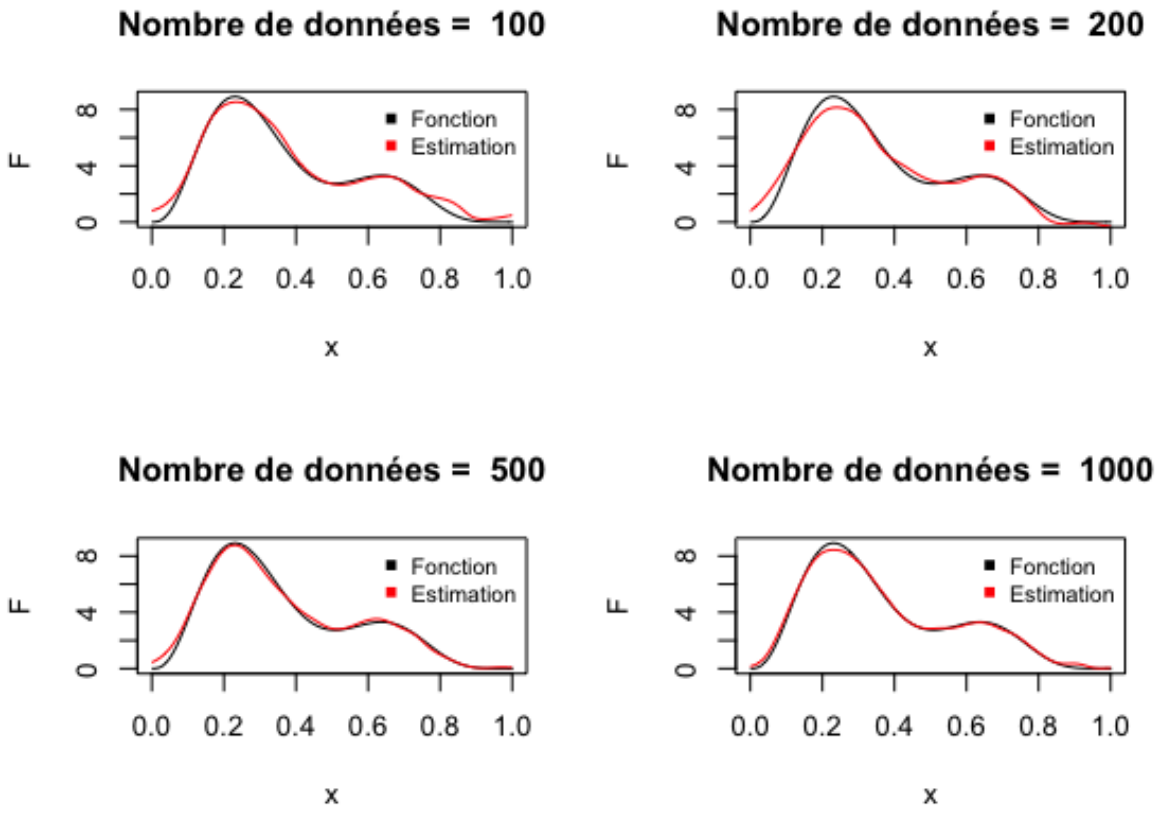
- $f(x) = 0,2x^{11} * (10 * (1-x))^6 + 10 * (10x)^3(1-x)^{10}$
- ϵ suit une loi normale N(0, 1)

Nous allons procéder à la même étude que précedemment.On veut visualiser l'influence d'une fenêtre de la forme $\widehat{\sigma_X} n^{-\alpha}$ où $\alpha \in [0;1]$ et $\widehat{\sigma_X}$ est l'estimation de l'écart type de X. On fixe n à 1000. Les résultats sont représentés par les graphiques ci-dessous :



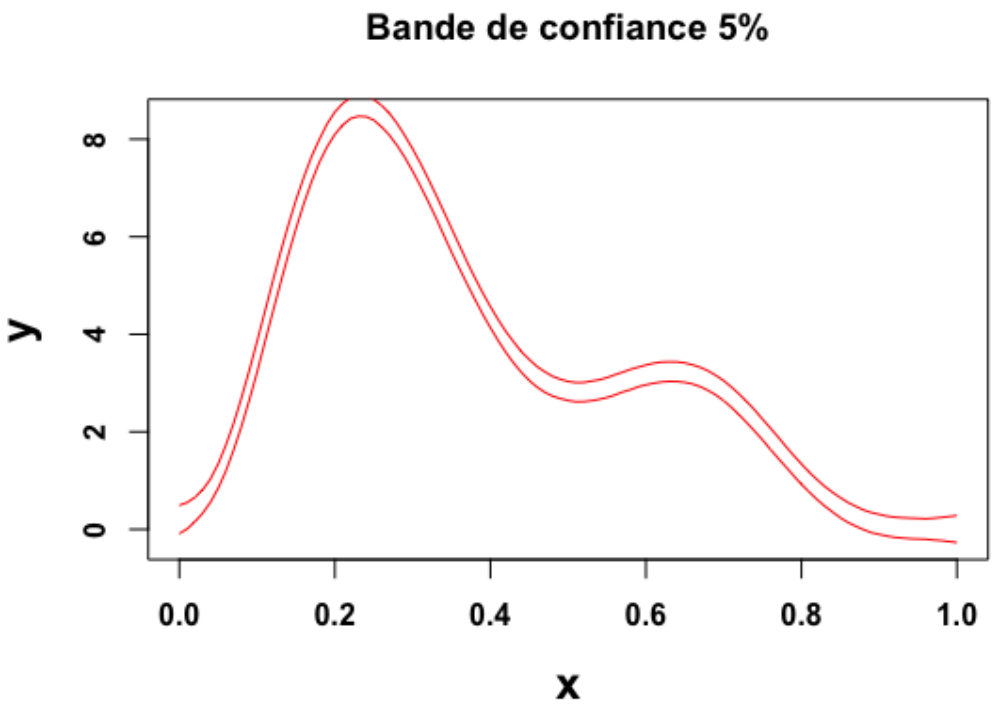
De même que précemment, on remarque que pour un α plus petit que 0.2, l'estimation est lisse mais biaisée. Inversement, pour un α plus grand que 0.2, l'estimation ne semble plus biaisée mais n'est pas lisse. Par conséquent, le bon compromis entre biais et variance s'opère pour un α de 0.2.

On veut maintenant étudier la qualité de l'estimation de la fonction f par l'estimateur de Nadaraya-Watson en fonction de la taille n de l'échantillon. On prend n=100, 200, 500 et 1000. On utilise un noyau gaussien. Cependant, dans ce cas le logiciel R ne trouve pas de fenêtre optimale par défaut. Nous fixons alors α à 0.2. Les résultats sont les suivants :

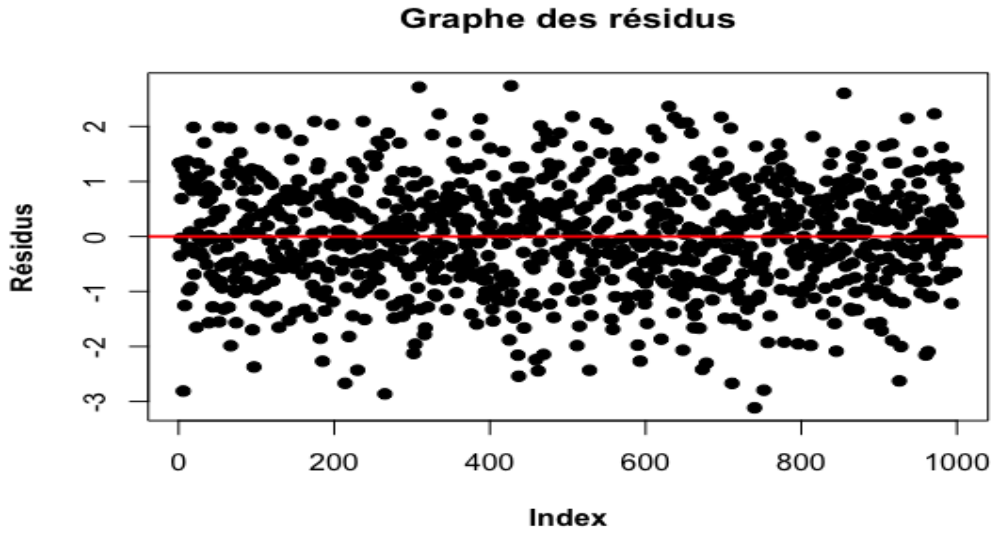


Comme avec la loi normale, plus la taille de l'echantillon est grande, meilleure est l'estimation. En effet l'estimation devient satisfaisante à partir de 500 données, en deça de cette valeur l'estimation diffère beaucoup de la fonction f.

On va maintenant utiliser la méthode de Monte-Carlo pour obtenir une bande de confiance de l'esimation de f. Nous utilisons la méthode des percentiles simples et effectuons 2000 simulations pour déterminer en toute valeur de t de notre intervalle d'étude l'intervalle de confiance pour f(t). On présente le résultat sur le graphique ci-dessous :



La bande est resserrée. Elle semble s'élargir légèrement lorsqu'on observe un changement de variation. Cependant on peut conclure une nouvelle fois que l'estimation de f est très bonne. Afin d'apprécier la qualité du modèle estimé on analyse maintenant les résidus. Ces derniers donnent le graphe suivant :



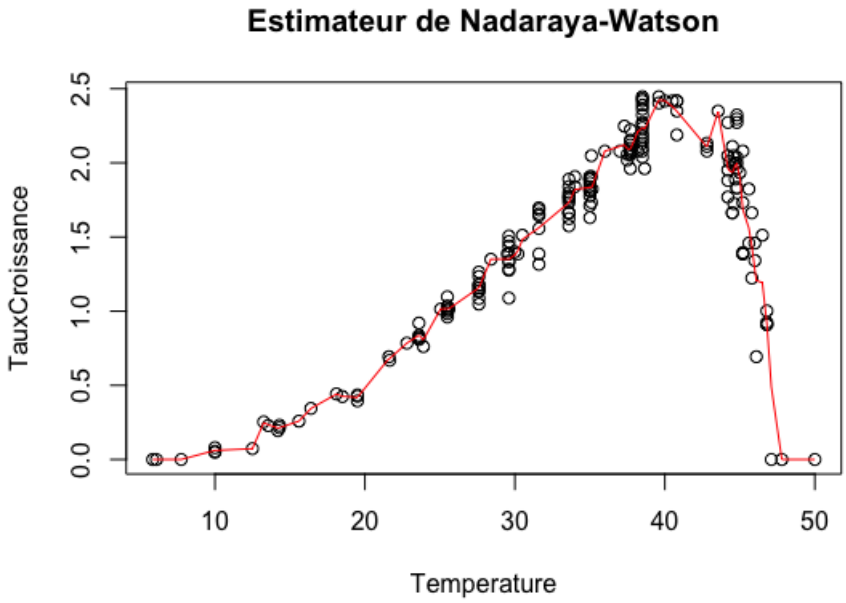
Les résidus sont assez élevés mais répartis équitablement autour de 0. On constate que l'hypothèse d'homocédasticité semble validée. De plus le test de Shapiro-Wilk donne une p-value de 0,60. On accepte donc au risque 5% le caractère gaussien des résidus. Enfin la moyenne quadratique des résidus est de 0.96, soit très proche de l'écart-type de ϵ qui est de 1. Ces éléments nous permettent de conclure que l'estimateur de Nadaraya-Watson est de bonne qualité.

Suite à ces deux études par simulations, nous constatons qu'avec un noyau gaussien l'estimateur de Nadaraya-Watson donne une très bonne estimation d'une fonction f dès n égale à 500. Mais également que le choix de la fenêtre a un véritable impact sur cet estimateur afin de vérifier le compromis biais-variance. Dans les deux cas un α proche de 0.2 apparait satisfaisant.

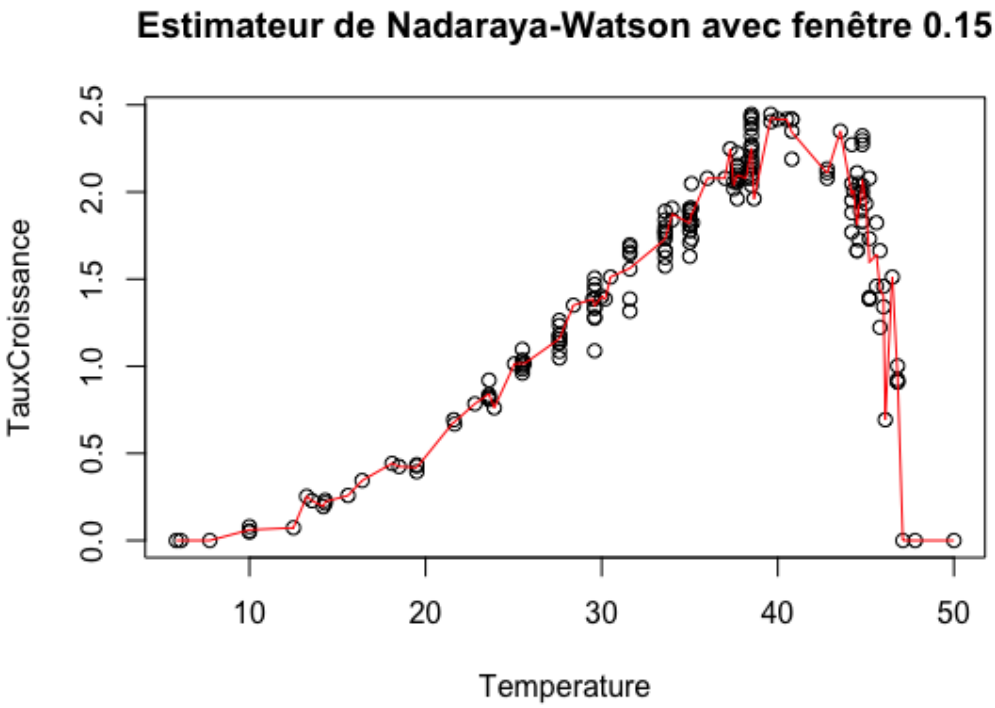
2 Deuxième partie : Expérimentation sur un jeu de données réelles

Dans cette seconde partie nous procédons à une nouvelle étude des données "Barber" utilisées lors de notre précédent TP. Cette fois nous souhaitons modéliser le taux de croissance de la bactérie à différentes températures à l'aide d'un modèle de régression non linéaire en la température.

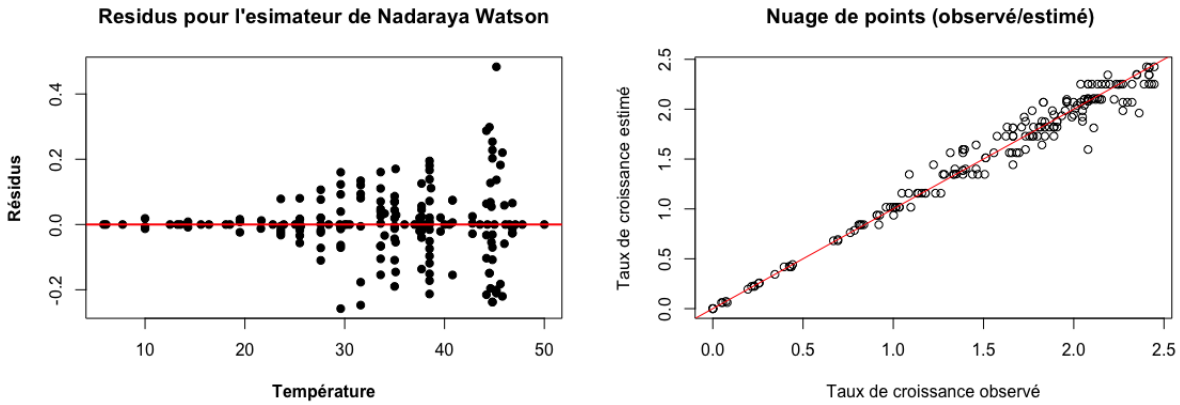
Dans un premier temps nous estimons la fonction de régression à l'aide de l'estimateur de Nadaraya-Watson en laissant le logiciel R fixer la fenêtre. Nous utilisons les données de l'échantillon d'apprentissage. Nous obtenons le graphe suivant :



Nous observons que le nuage de points est très bien ajusté par l'estimateur de Nadaraya-Watson, mais la courbe n'est pas lisse. La somme des carrés des erreurs est de 3.01 et est plus petite que dans le cas de l'ajustement non linéaire paramétrique.
On veut minimiser la somme des carrés des erreurs. Comme on a vu qu'il est important de bien choisir la fenêtre, on modifie la valeur de la fenêtre et on la choisit de la forme suivante : sd(T)$n^{-0.2}$ où sd(T) désigne l'écart-type des températures observées. La valeur de cette fenêtre est 3.44. On va calculer alors la valeur de la fenêtre optimale par validation croisée. On se donne l'intervalle]0.0,8.0]. On trouve la valeur 0.15. Nous estimons alors de nouveau la fonction de régression à l'aide de l'estimateur de Nadaraya-Watson en prenant comme valeur de fenêtre 0.15. On obtient le graphique suivant :



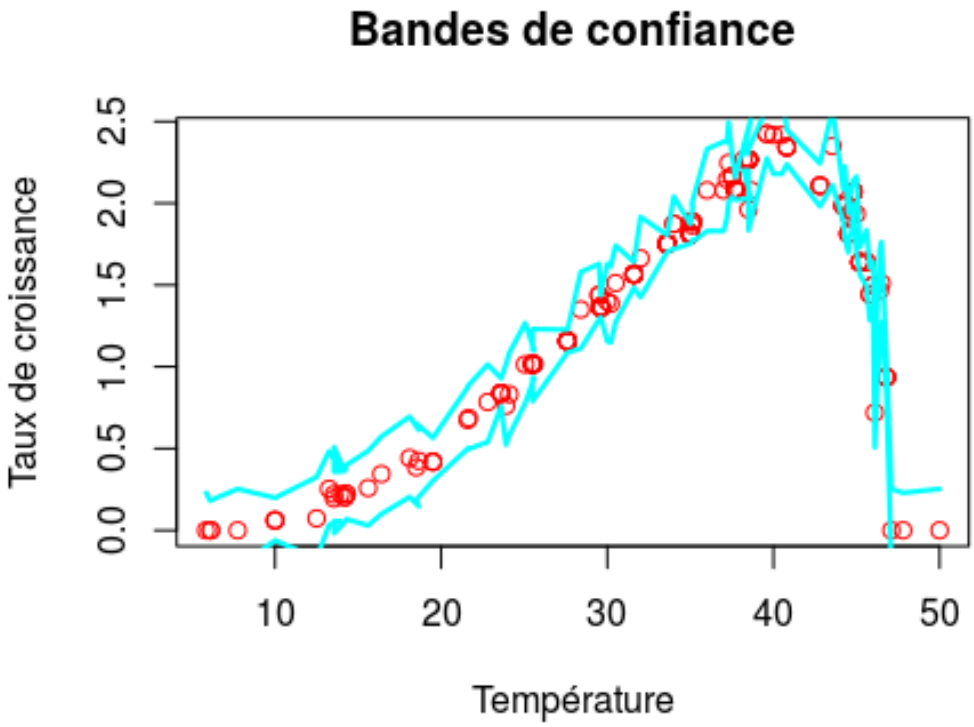
On observe que le nuage de points est encore mieux ajusté ce qui implique que la courbe est encore moins lisse. Dans ce cas la somme des carrés des résidus vaut 2,08. On étudie le nuage de points observés-estimés ainsi que le graphe des résidus ci-dessous :



Si l'on compare ces deux graphiques à ceux obtenus lors du TP précédent on peut dire que le nuage de points observé/estimé est meilleur. En effet on ne détecte aucune sur ou sous-estimation, contrairement aux résultats obtenus dans le TP 4. Concernant

les résidus, ils sont plus faibles mais l'hypothèse d'homocédasticité n'est toujours pas vérifiée. On peut conclure en disant que cette approche semble plus performante lorsque l'on recherche le meilleur ajustement possible. Mais si on souhaite faire de la prévision, elle semble moins appropriée car l'estimation n'est pas lisse.

Pour finir, on souhaite construire une bande de confiance à 95% pour la fonction estimée. On s'aide d'une méthode bootstrap basée sur les résidus. On obtient le résultat suivant :



Comme on pouvait s'y attendre, les bandes ne sont pas du tout lisses mais sont resserées. Elles semblent cependant un peu moins resserées que les bandes de confiance trouvées lors du TP4.