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

Régression Non Linéaire avec Applications sous R

Rapport de TP n°6

Titre : *Estimation par backfitting dans les modèles additifs non linéaires*

L'objectif de ce TP est d'expérimenter l'algorithme du backfitting dans le cadre des modèles additifs non linéaires. On cherche à expliquer le maximum de la concentration en ozone du jour (en $\mu g/m^3$) à partir de trois variables explicatives. On donne les trois variables suivantes :

- temp : le maximum de température du jour en $°C$
- vent : le maximum de vitesse du vent entre 14h et 18h en m/s
- mozon : la moyenne spatiale de la concentration en ozone sur la région en $\mu g/m^3$

Pour cela, on utilise un modèle additif non linéaire de la forme :

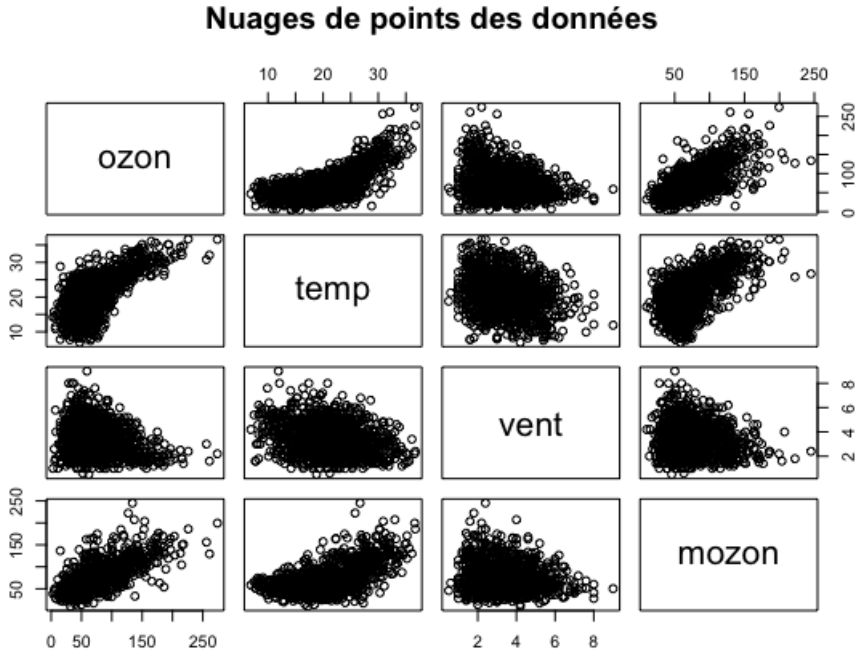
$$O = f_1(temp) + f_2(vent) + f_3(mozon) + \mu + \epsilon$$

Avec :

- μ : Une constante
- ϵ : Un terme d'erreur satisfaisant où $E[\epsilon] = 0$
- $E[f_j(X_j)] = 0, \forall j \in \{1, 2, 3\}$

Comme pour les TP précedents, nous séparons les données en deux échantillons. Un échantillon d'apprentissage qui comprend 80% des données et servira à mettre au point le modèle et un échantillon test, qui sera composé des 20% données restantes et on l'utilisera pour tester la capacité du modèle à faire des prédictions.

Comme toujours avec ce type de données, nous allons commencer par étudier la possible corrélation entre les variables disponibles. On trace les graphes suivants :



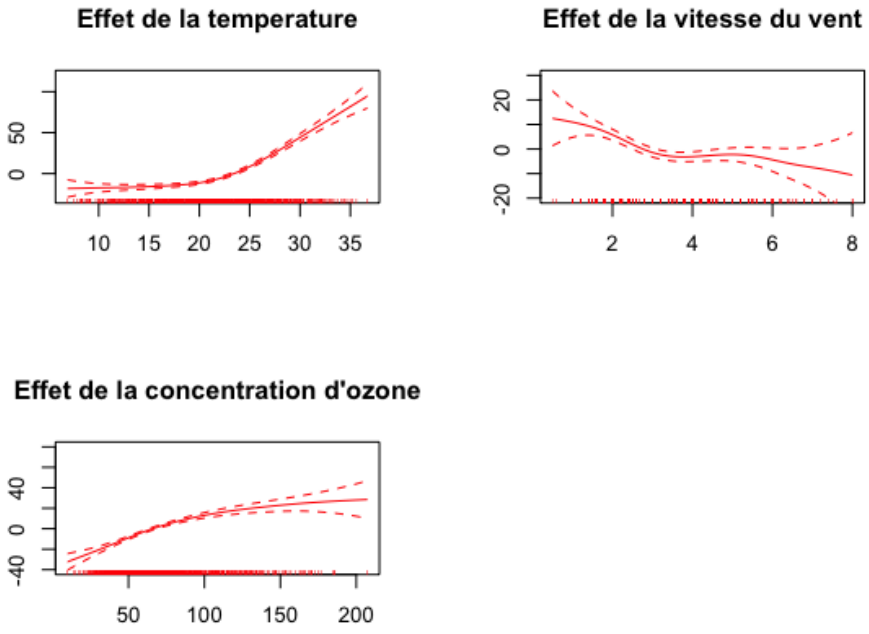
Aucun nuage de points ne forme une droite. De plus, la matrice de corrélation donne des valeurs très différentes de 1. On peut alors conclure qu'il n'existe pas de forte corrélation entre les deux variables.

On modèlise maintenant les données à l'aide d'un modèle additif non linéaire. On obtient le tableau résultat suivant :

Coefficient	edf	Ref .df	F	p-value
s(temp)	3.985	4.970	93.871	$< 2{*}10^{-16} * * *$
s(vent)	4.023	4.994	7.358	$9.3{*}10^{-7} * * *$
s(mozon)	3.000	3.803	50.057	$< 2{*}10^{-16} * * *$

De plus, la constante μ est estimée à 72.206. Sa p-value associée est inférieure à $2{*}10^{-16}$.

Les p-values associées à chaque variable sont toutes inférieures à 5%. On accepte donc au risque 5% la significativité de l'effet des variables. De plus, les estimations des degrés de liberté des variables sont toutes largement supérieures à 1. On peut donc dire que les variables ne sont pas liées linéairement à la concentration en ozone. Tout cela valide l'utilisation d'un modèle non linéaire plutôt qu'un modèle linéaire. De plus, nous pouvous estimer les fonctions f_1, f_2 et f_3 afin d'expliciter les effets des différentes variables explicatives sur le taux d'Ozone. Les graphiques ci-dessous montrent les estimations de ces fonctions :

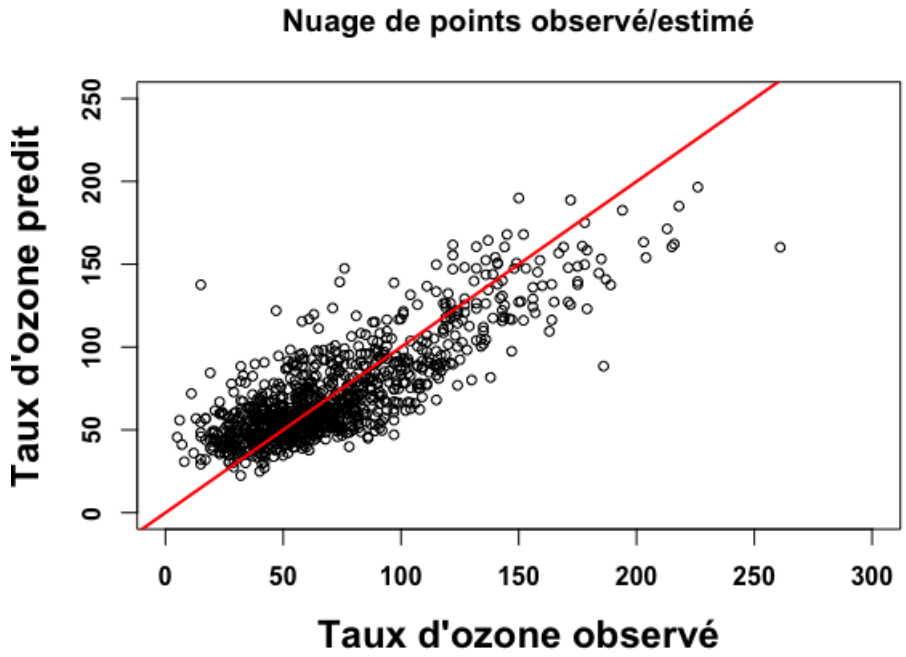


Les 3 variables ont des effets différents :

- Température : elle n'a pas beaucoup d'effet jusqu'à $20°C$. Après $20°C$, plus la température est élevée, plus son effet sur la concentration est important.
- Vent : l'effet du vent est le moins important des trois variables. Pour une vitesse de vent compris entre 3 et 6 m/s, l'effet du vent est négligeable. Autrement un vent très faible a un effet positif sur la concentration en ozone tandis qu'un effet négatif est engendré par un vent plus puissant.
- Mozon : la concentration de Mozon croit avec son effet sur la concentration d'Ozone. Pour une concentration en Mozon compris entre 0 et 120 $\mu g/m^3$, l'effet croit de manière importante allant de -40 jusqu'à 35. Au dessus d'une concentration de 100 $\mu g/m^3$ l'effet continue d'augmenter très légèrement jusqu'à 40.

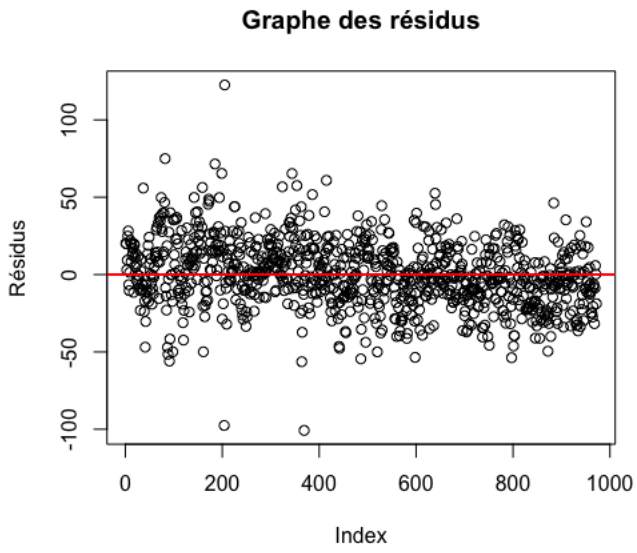
On remarque que la bande de confiance est plus resserée quand il y a plus d'observations. Elle est plus large pour les valeurs très faibles et très grandes de vent et mozon, là où l'on observe peu de données.

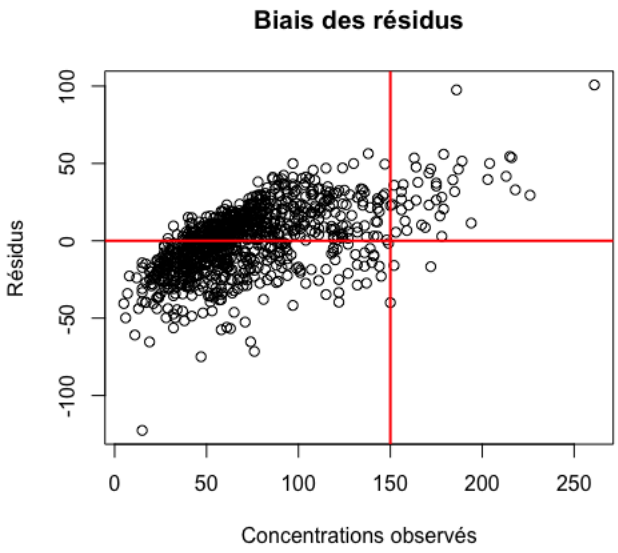
On souhaite maintenant étudier la qualité du modèle. Or le taux de variance expliquée est de 67%. Cette valeur laisse présager des erreurs dans le modèle. Nous étudions le nuage de points (observés, estimés) pour rendre compte de ces possibles erreurs. On obtient le graphique suivant :



On remarque une importante sous-estimation pour une grande concentration en ozone, supérieure à 150 $\mu g/m^3$. Dans le cadre de notre problème, ce résultat est problématique

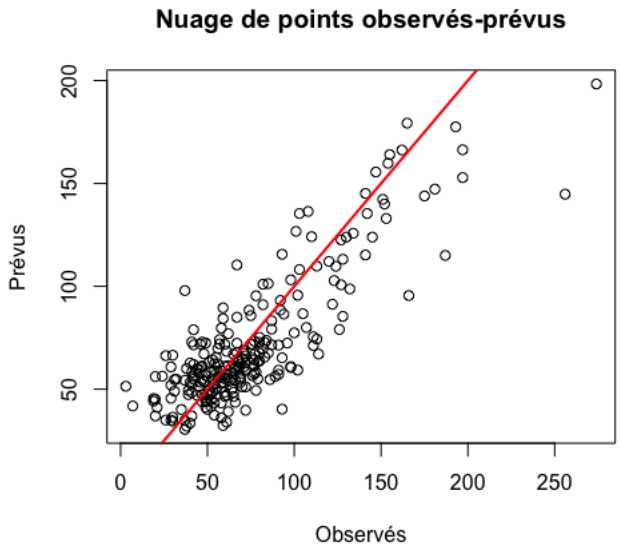
puisque ce sont justement les grandes concentrations qui nous intéressent. En effet, le but de cette étude est de prévoir les pics de pollution pour réagir en conséquence. Mais notre modèle sous-estime ces pics de pollution. Regardons maintenant le graphe des résidus et le graphique représentant le biais des résidus :

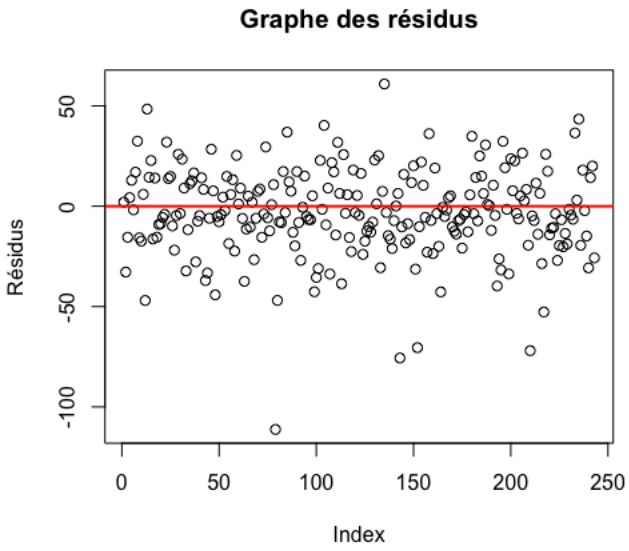




La remarque précédente est bien vérifiée par l'étude du biais des résidus. Comme on peut le voir sur le graphique, on retrouve bien la sous-estimation de la concentration en ozone dès lors que cette concentration dépasse 150 $\mu g/m^3$. De plus, on observe sur le graphe des résidus que les résidus sont très élevés. En effet, on atteint de grandes valeurs comprises entre -100 et 100. Mais visuellement ils semblent répartis équiblement autour de 0 et l'hyphotèse d'homocédasticité parait vérifiée. Par ailleurs, la valeur du MAE est de 16.7 et celle du RMSE est de 21.6, valeurs très élevées.
Un test de Shapiro-Wilk rejette la normalité des résidus au risque 5%. On pouvait s'y attendre étand donnée la sous-estimation des fortes concentrations qui implique une asymétrie dans la distribution des résidus.

On veut maintenant étudier les performances du modèle sur l'échantillon test. Nous traçons alors le nuage de points (observés, prévus) et le graphe des résidus suivant :





Comme c'était le cas pour l'échantillon d'apprentissage, on peut observer une sous-estimation des fortes concentrations en ozone, cas de forte pollution. Le graphe des résidus nous permet de même de dire que l'hypothèse d'homocésticité parait vérifiée car les résidus sont répartis équitablement autour de 0. Les résidus sont aussi élevés.

On peut classer la concentration en ozone en trois intervalles :

- Niveau 0 : [0,130] $\mu g/m^3$
- Niveau 1 : [130,180] $\mu g/m^3$
- Niveau 2 : [180,240] $\mu g/m^3$

Dans le cas de la prediction de l'échantillon de test, on a le tableau d'alertes suivant :

	Niveau 0	Niveau 1	Niveau 2
Niveau 0	217	2	0
Niveau 1	6	11	0
Niveau 2	1	5	1

On en déduit le tableau de dépassement :

	Dép. non prévus	Dép. prévus
Dép. non réalisés	217	2
Dép. réalisés	7	17

Une nouvelle fois, on remarque une tendance du modèle à ne pas prévoir les hautes concentrations. On remarque qu'il y a 2 faux positifs contre 7 faux négatifs. Or le but d'un modèle est d'avoir des nombres de faux positifs et faux négatifs similaires.
On regarde les indicateurs classiques :

- POD : 0.71%
- FAR : 0.11%
- TS : 0.65%
- SI : 0.7%

Les 4 taux calculés sont corrects. On s'intéresse au taux "d'oubli" du modèle, c'est-à-dire les cas où le modèle n'a pas su détecter un dépassement de la concentration en ozone. Ce taux est de 24%, ce qui est trop important.
On calcule enfin 4 indices d'erreurs en prevision :

- R : 0.85
- EV : 0.71
- MAE : 16.59
- RMSE : 22.09

Le coefficient de corréalation entre observés et prévus est assez élevé mais la moyenne de l'erreur absolue et l'erreur quadratique moyenne sont elles aussi très élevées. Le RMSE et le MAE sont à 0.5 près identiques au RMSE et MAE obtenus avec l'échantillon d'apprentissage. Cela nous permet de dire que les performances du modèle ne sont pas dégradées lorsque l'on passe de l'échantillon d'apprentissage à l'échantillon test.

En conclusion, à la suite de l'analyse du modèle non linéaire, on peut dire qu'il ne semble pas très performant. En effet, les indicateurs ne sont pas excellents. De plus, le modèle sous-estime grandement les grandes concentrations en ozone. Ceci est problématique puisque ce modèle a pour but de prévoir les épisodes de forte concentration en ozone qui sont associés à un pic de pollution...

On veut maintenant comparer le modèle non linéaire avec un modèle linéaire. On suppose donc que les concentrations de l'échantillons d'apprentissage y_j sont les réalisations de variables aléatoires Y_j liées aux x_{ji} avec $i \in \{1, 2, 3\}$ et $j \in \{1, ..., 942\}$ par la relation :

$$Y_j = \mu + \sum_{i=1}^{3} \alpha_i x_{j,i} + \epsilon_j$$

où :

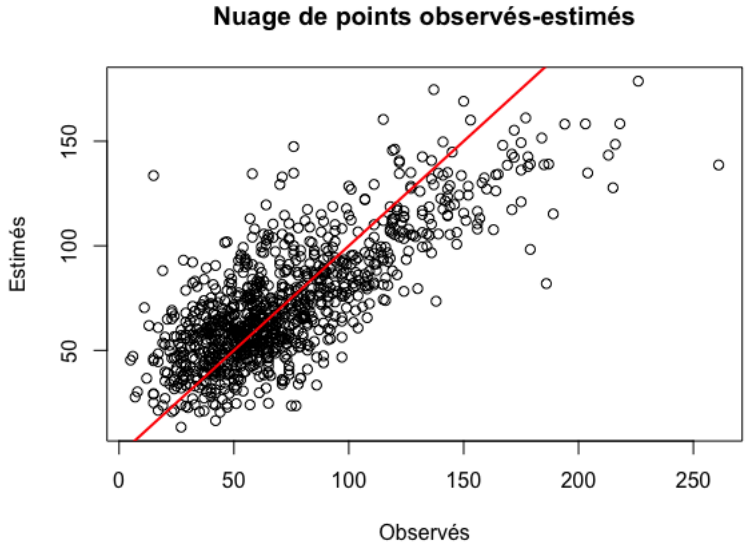
- μ et α_i sont des paramètres réels inconnus,
- les erreurs (ϵ_j) sont des variables aléatoires que l'on suppose indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$

On calcule les coefficients estimés du modèle :

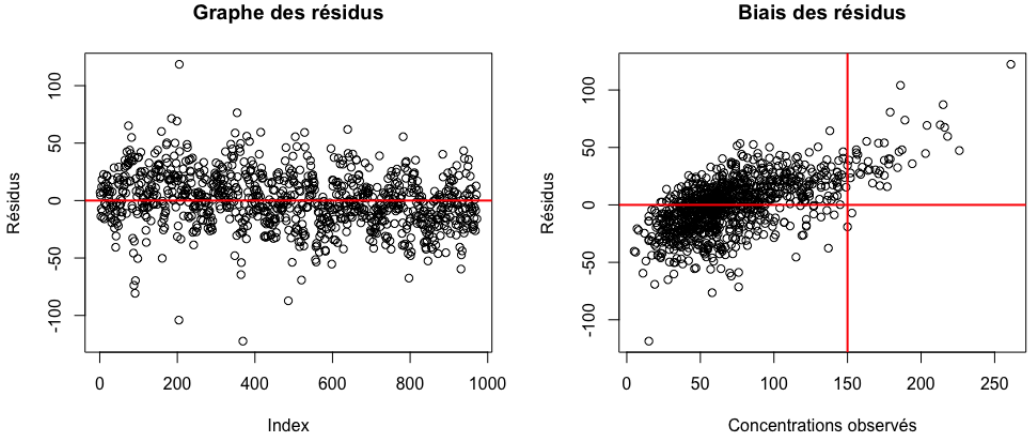
- μ = -15.04
- α_1 = 2.84

- α_2 = -2.87
- α_3 = 0.52

De plus, les p-values de toutes les variables sont inférieures à 5%, ce qui valide leur significativité. La part de variance expliquée du modèle est de 59,8%, inférieure à celle obtenue avec le modèle non linéaire.
On trace maintenant le nuage de points (observés, estimés) :

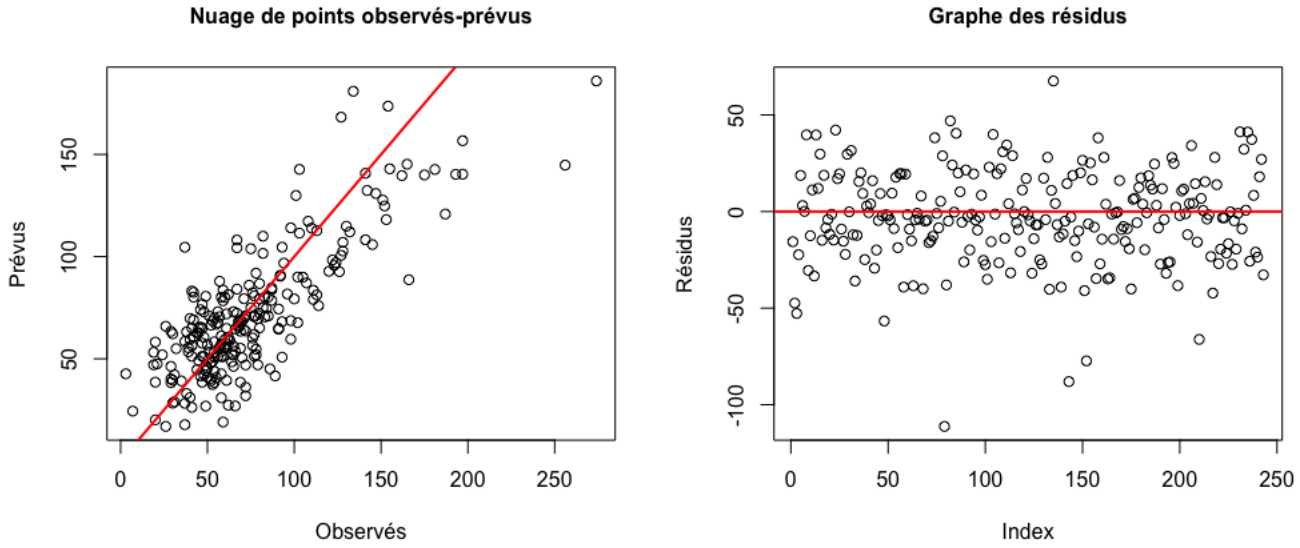


A priori, on retrouve le problème de sous-estimation du premier modèle. On étudie à nouveau les résidus en traçant le graphe des résidus et le graphe représentant le biais des résidus.



Il est difficile de comparer les deux modèles avec ces graphiques qui semblent similaires. En effet tout comme avec le graphe du modèle non linéaire, on remarque une importante sous-estimation à partir d'une concentration supérieure à $150\mu g/m^3$. D'autre part les résidus sont élevés bien qu'ils soient identiquement répartis autour de 0 et l'hypothèse d'homocédasticité est vérifiée. La valeur du RMSE est de 23.88 et celle du MAE est de 18.36. Ces deux valeurs nous permettent de dire que le modèle de regression linéaire semble légèrement moins performant que le modèle de régression non linéaire. Enfin un test de Shapiro-Wilk rejette la normalité des résidus au risque 5%.

On veut maintenant étudier les performances du modèle sur l'échantillon test. On prédit les concentrations en ozone de l'échantillon test à l'aide du modèle linéaire. Nous traçons alors le nuage de points (observés, prévus) et le graphe des résidus suivant :



Comme c'était le cas pour l'échantillon d'apprentissage, on peut observer une sous-estimation des fortes concentrations en ozone, cas de forte pollution. Le graphe des résidus nous permet de même de dire que l'hypothèse d'homocésticité parait vérifiée car les résidus sont répartis équitablement autour de 0. Les résidus sont aussi élevés.

On peut ensuite afficher le tableau d'alerte associé au modèle linéaire :

	Niveau 0	Niveau 1	Niveau 2
Niveau 0	217	2	0
Niveau 1	8	8	1
Niveau 2	1	5	1

Et on en déduit le tableau de dépassement :

	Dép. non prévus	Dép. prévus
Dép. non réalisés	217	2
Dép. réalisés	9	15

On peut voir que les résultats restent presque semblables. On observe seulement deux cas supplémentaires de dépassements non prévus réalisés .
On calcule alors les indicateurs classiques :

- POD : 0.62%
- FAR : 0.12%
- TS : 0.58%
- SI : 0.62%

En comparaison avec les taux précédents, tous les indices sont moins bons. On calcule de nouveau le taux "d'oubli" qui est de 38%. Cette valeur est encore une fois plus élevée que celle calculée avec le modèle précedent.
Les indices d'erreurs en prevision sont les suivants :

- R : 0.81%
- EV : 0.65%
- MAE : 18.28%
- RMSE : 24.12%

Le RMSE et le MAE sont à 0.5 près identiques au RMSE et MAE obtenus avec l'échantillon d'apprentissage. Cela nous permet de dire que comme pour le modèle de regression non linéaire, les performances du modèle ne sont pas dégradées lorsque l'on passe de l'échantillon d'apprentissage à l'échantillon test. En comparant ces indices avec ceux du modèle de régression non linéaire étudié précedemment, on s'aperçoit qu'ils sont, à nouveau, tous un peu moins bons

Finalement, toutes les comparaisons indiquent que le modèle non linéaire n'approxime que légèrement mieux la concentration en ozone que le modèle linéaire. Mais les deux modèles présentent un problème de taille : la sous-estimation de la concentration en ozone quand celle-ci est élevée. Ainsi il est très compliqué de faire de la prévision avec l'un des deux modèles.