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

Régression Non Linéaire avec `R`

Rapport de TP n°1

Titre : *Prise en main du Logiciel* ***`R`***

1 Introduction

L'objectif de ce premier TP est d'expérimenter la sélection de variables ascendantes et descendantes, tout en s'initiant à l'utilisation du logiciel R. C'est dans cette optique que nous cherchons à connaître l'influence de certaines variables sur la densité de peuplement d'un parasite : la chenille processionnaire du pin. Pour ce faire, le contexte suivant est donné :
On dispose d'un échantillon de n = 32 parcelles forestières d'une surface de 10 hectares. Chaque parcelle est alors échantillonnée en placettes de 5 ares et on a calculé les moyennes (sur ces placettes) des quantités suivantes :

- X1 : l'altitude en mètres
- X2 : la pente en degrés
- X3 : le nombre de pins dans la placette
- X4 : la hauteur en mètres de l'arbre échantillonné dans la placette
- X5 : le diamètre de cet arbre
- X6 : la note de densité de peuplement
- X7 : l'orientation de la placette (de 1=sud 2=autre)
- X8 : la hauteur en mètres des arbres dominants
- X9 : le nombre de strates de végétation
- X10 : le mélange de population (de 1 : mélangé à 2 : non mélangé)

2 Etude statistique

2.1 Corrélation

Avant d'introduire quelconque modèle, il est nécessaire de s'interesser aux possibles corrélations entre les variables explicatives du jeu de données, pour ne garder que des variables explicatives non corrélées.
Dans un premier temps, on trace le graphique représentant les données de chaque variable en fonction des autres variables :

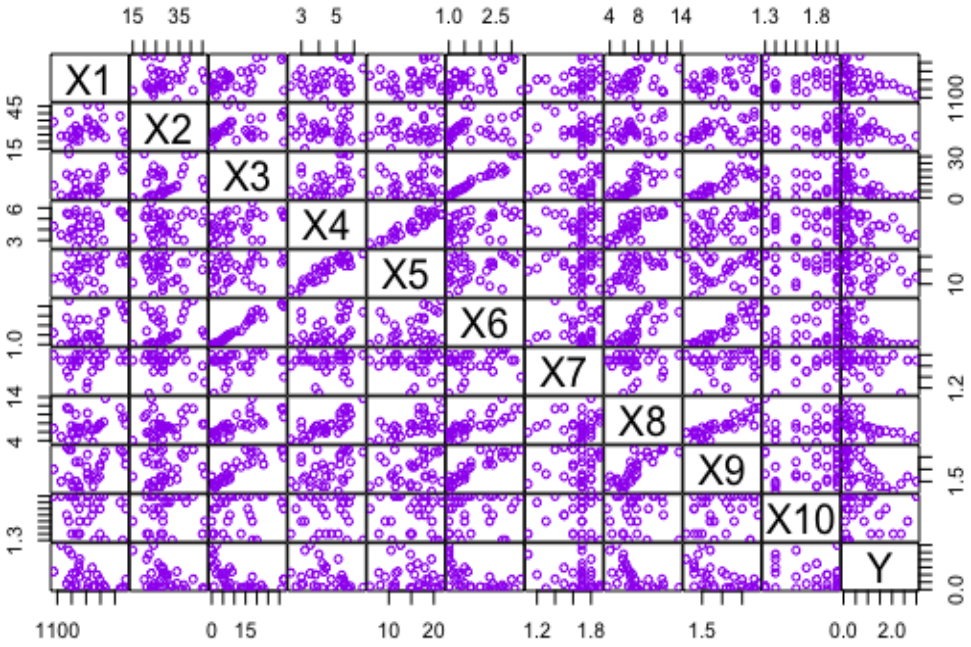

FIGURE 1 – Graphique des données X_j en fonction de X_i, pour i allant de 1 à 10 et i différent de j

Une forte corrélation entre deux variables se présente sous la forme d'une droite alors qu'une non correlation se presentera sous la forme d'un nuage de points dispersés. Ainsi nous pouvons voir à l'oeil nu sur la figure ci dessus que les variables X3-X6, X4-X5, X6-X9 et X8-X9 semblent corrélées deux a deux. Via le logiciel R avec la fonction "cor", on crée alors la matrice de corrélation des données et on a :

- Corrélation(X3,X6) = 0.979
- Corrélation(X4,X5) = 0.905
- Corrélation(X6,X9) = 0.902
- Corrélation(X8,X9) = 0.831

Ces résultats indiquent un lien très important entre les variables X3 et X6, c'est-à-dire entre le nombre de pins dans une placette et la note de densité de peuplement. En effet, puisque toutes les placettes ont la même taille, la densité dépend uniquement du nombre de pins dans la parcelle.
Les autres coefficients de corrélations, bien qu'ils soient importants, ne permettent pas d'écarter une variable compte tenu le faible nombre de données à notre disposition.

Ainsi, après cette analyse, nous allons envisager de nous séparer de la variable X3 ou X6. Pour valider ce choix, nous utilisons le VIF (Variance Inflation Factor) pour étudier la multi-colinéarité. La fonction VIF sur R nous donne les résultats suivants :

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Valeur	1.9	1.2	42.8	16.3	9.7	61.5	1.7	13.2	9.6	1.7

Une valeur supérieure à 20 indique qu'une variable est fortement liée avec les autres variables. C'est le cas de X3 et X6. Puisque X6 est la variable pour laquelle l'indicateur VIF est le plus important, c'est cette variable qu'on préfère retirer du modèle par la suite. Nous avons réitérer l'étude de multicolinéarité avec l'indicateur VIF en enlevant la variable X6 pour vérifier que la valeur du VIF des neuf variables restantes est en dessous de 20.

2.2 Modèle statistique

Dans cette partie on souhaite expliquer la variable d'intérêt Y, la densité de peuplement, par une liaison linéaire à partir des différentes variables explicatives. On introduit donc le modèle linéaire gaussien multiple, c'est à dire que l'on suppose que les données (y_j) sont les réalisations des 32 variables aléatoires (Y_j) liées aux (x_{ji}) avec $i \in \{1; 2; ...; 10\}$ et $j \in \{1; 2; ...; 32\}$ par la relation :

$$Y_j = \mu + \sum_{i=1}^{10} \alpha_i x_{j,i} + \epsilon_j$$

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 $\mathcal{N}(0, \sigma^2)$

L'étude précédente nous a permir d'exclure la variable X6. Ainsi nous reprenons le modèle ci-dessus mais avec $i \in \{1; 2; ...; 10\} - \{6\}$.

On calcule alors les coefficients estimés du modèle :

- μ = 8.484772785
- α_1 = -0.003211428
- α_2 = -0.042409502
- α_3 = 0.025318248
- α_4 = -0.474230295
- α_5 = 0.126756950
- α_7 = -0.065284453
- α_8 = -0.064355931
- α_9 = -0.444014909
- α_{10} = -0.501254601

On va maintenant étudier la qualité de l'ajustement linéaire des données. On effectue tout d'abord un test de non regréssion pour savoir si les neuf variables explicatives ont une réelle influcence sur Y.
On teste donc l'hypothèse nulle H_0 :" $\alpha_1 = \alpha_2 = ... = \alpha_{10} = 0$"
contre l'hypothèse alternative H_1 :" $\exists$i$\in \{1; 2; ...; 10\} - \{6\}$ tel que $\alpha_i \neq 0$" .

Le logicel R nous donne une p-value de $6,066.10^{-4}$, soit une p-value inférieure à 5%.

On rejette donc l'hypothèse nulle au risque 5% et on en déduit qu'au moins une des variables explicatives a une influence significative sur le nombre de nids par arbre.
De plus, la part de variance expliquée par le modèle de 68,81% et celle ajustée de 56,05%, nous montre qu'avec neuf variables explicatives, on peut expliquer une grand part du modèle.

Observons maintenant sur le graphique ci dessous le nombre moyen de nids par arbre estimé par le modèle en fonction du nombre observé :

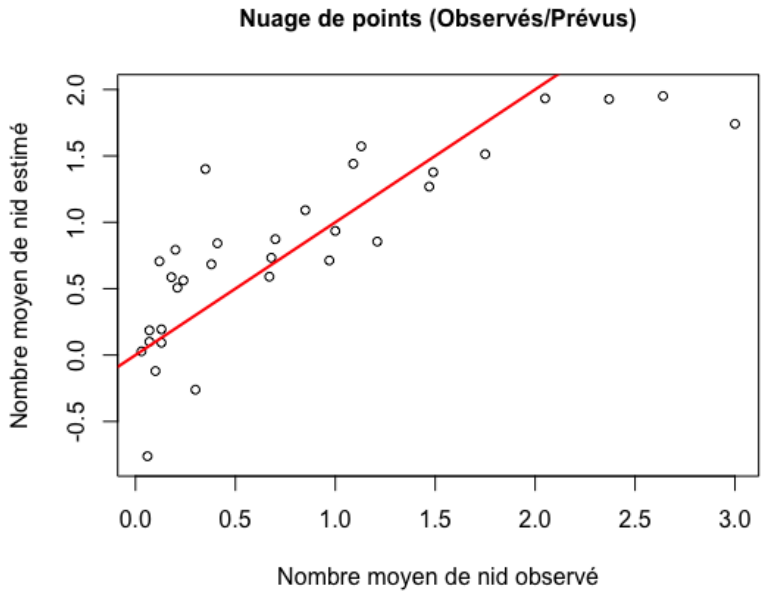

FIGURE 2 – Nuage de points des valeurs estimées en fonction des valeurs observées.

On remarque qu'une tendance linéaire apparait clairement. On peut aussi observer que le modèle a tendance à estimer à la hausse lorsque le nombre observé de nids est petit, et inversement à estimer à la baisse lorsque le nombre observé de nids est grand.

Nous étudions maintenant le graphe des résidus. Avec le logiciel R, nous traçons le graphe des résidus et le graphe des résidus studentisés.

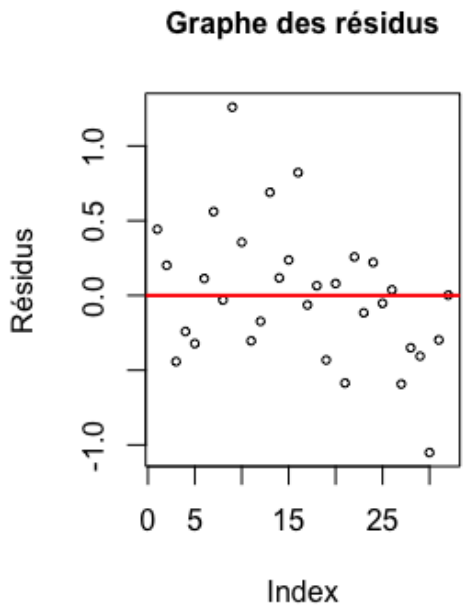

FIGURE 3 – Graphe des résidus

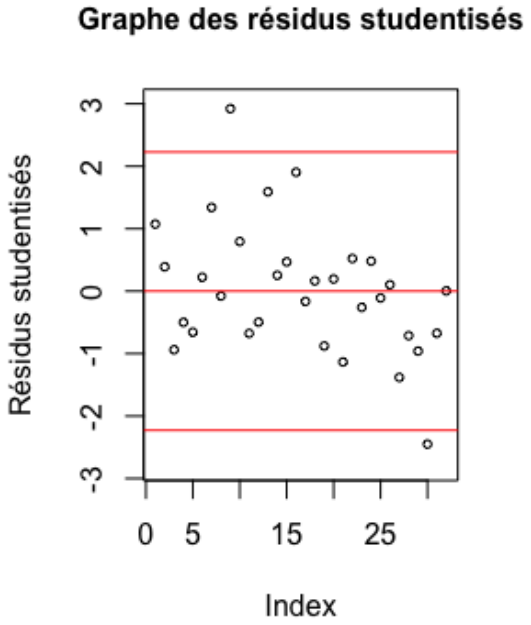

FIGURE 4 – Graphe des résidus studentisés (seuils à 95%)

On peut observer sur le graphe des résidus que les points sont répartis de manière équitable autour de 0 et les valeurs des résidus ne sont pas trop élevées. Nous n'observons pas d'augmentation significative de l'amplitude des résidus en fonction du nombre moyen de nids par abre, l'hypothèse d'homoscédacité est donc vérifiée. Le graphe des résidus studentisés permet de s'interroger sur le caractère aberrant ou non des données du jeu 9 et 30. En effet, on observe que les résidus pour le jeu 9 et 30 sont en dehors de bande de confiance à 95%.

Nous testons maintenant la normalité des résidus par le biais du test de Shapiro-Wilk. Nous obtenons une p-value de 0,6996 supérieure à 5%. Ainsi on accepte au risque de 5% l'hypothèse que les résidus suivent une loi normale. On peut vérifier ce résultat visuellement avec l'histogramme des résidus ci-dessous où on observe bien une forme gaussienne dans la répartition des résidus.

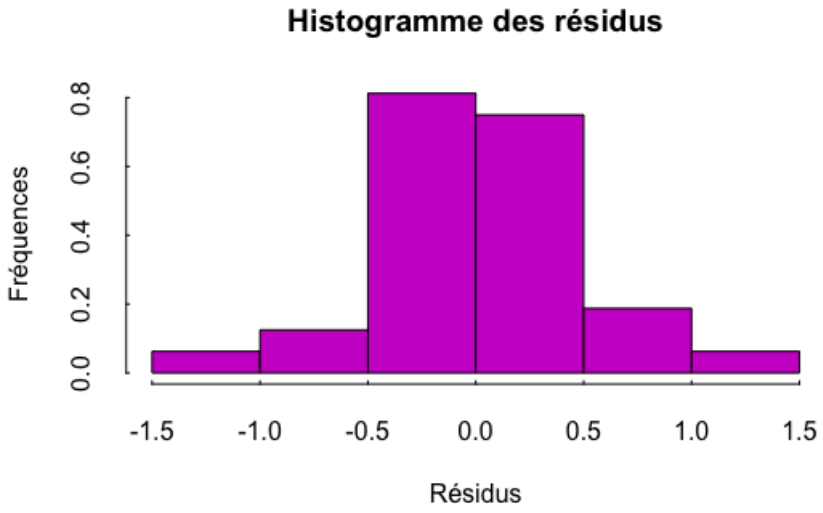

FIGURE 5 – Histogramme des résidus

Enfin en effectuant le test de Durbin-Watson, le logiciel R donne une p-value égale à 0.168 et donc supérieure à 5%. Ceci nous permet d'accepter au risque 5% l'hypothèse de non corrélation des résidus.

En conclusion, au vu de ces dernières observations nous pouvons dire que l'ajustement est de bonne qualité.

2.3 Forward regression et Backward regression

Bien que les résultats obtenus précedemment puissent être considérés satisfaisant, il est parfois intéressant de ne prendre en considération qu'une partie des variables disponibles. C'est pourquoi nous chechons maintenant à ne garder dans le modèle que les variables explicatives qui ont une influence réelle sur la variable à expliquer Y, tout en gardant un modèle efficace.
Pour cela, nous allons utiliser deux méthodes : la regression pas à pas ascendante (forward regression) et la regression pas à pas descendante (backward regression).

Backward regression

Dans cette méthode, on part du modèle complet (les 9 variables X1, X2, X3, X4, X5, X7, X8, X9 et X10) et à chaque étape on élimine la variable la moins significative au risque 5%. Pour ce faire on calcule les p-values associées à chaque variable et on cherche la plus grande de ces p-values qui est supérieure à 5%. Si elle existe, on retire alors cette variable du modèle. On réitère ce shéma jusqu'à ne plus pouvoir éliminer de variables. En appliquant cette méthode, on supprime successivement les variables suivantes :

- $X7$ de p-value 0.915
- $X8$ de p-value 0.630
- $X3$ de p-value 0.364
- $X10$ de p-value 0.294

- $X9$ de p-value 0.155

On arrive donc à un modèle gardant 4 variables explicatives : X1, X2, X4, X5. On peut alors expliquer 63% de la variance.

Forward regression

Le principe de la regression pas à pas ascendante est de sélectionner une à une la variable explicative la plus significative (p-value inférieure à 5% et la plus petite) puis de l'ajouter à la régression suivante. On s'arrête lorsqu'il n'y a plus de variable significative. Nous avons alors dans l'ordre garder les variables suivantes :

- $X9$ de p-value $4.53.10^{-4}$
- $X1$ (régression linéaire à 2 variables explicatives avec X9) de p-value $1.18.10^{-2}$
- $X2$ (régression linéaire à 3 variables explicatives avec X9 et X1) de p-value $1.23.10^{-2}$

Les p-values sont ensuite toutes supérieures à 5% en faisant une regression linéaire à 4 variables explicatives avec X9, X1 et X2 avec toutes les variables restantes. Cette méthode nous donne donc un modèle comportant les variables X1, X2, X9 et permet d'expliquer 58% de la variance.

Modèle simplifié

Nous avons vu qu'avec les deux méthodes précédentes, nous pouvons garder la moitié ou le tiers des variables explicatives que nous avions au début de notre étude, tout en conservant une grande partie de la variance expliquée. On effectue alors une régression avec les variables que nous avons conservées pour les deux méthodes, soient X1, X2, X4, X5 et X9. Nous obtenons alors 65,8% de variance expliquée. C'est seulement 3% de moins que la variance expliquée obtenue avec 9 variables explicatives, alors qu'ici, nous n'en avons conservé que 5.

Etudions maintenant la qualité de notre nouveau modèle. La figure ci dessous représente le nombre moyen de nids par arbre estimé par le modèle simplifié en fonction du nombre observé. Nous observons une nouvelle fois que le modèle que nous avons construit surestime le nombre de nids lorsque celui-ci (observé) est petit et inversement, le modèle sous estime ce nombre.

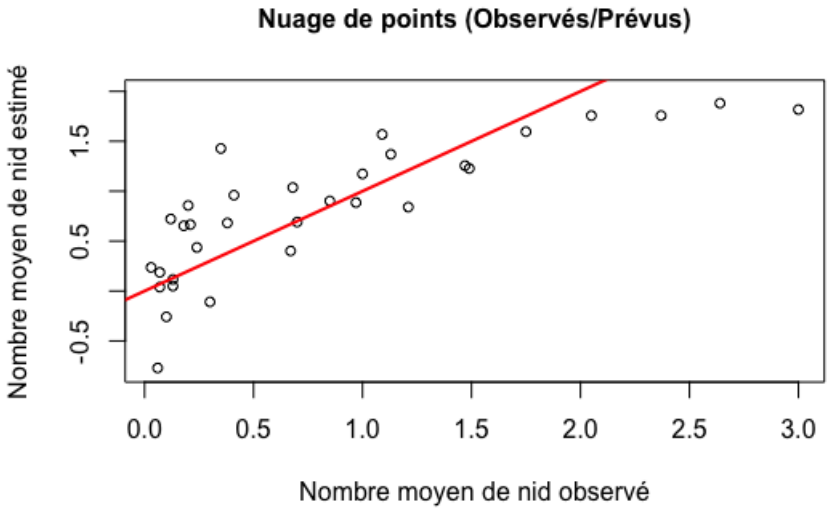

FIGURE 6 – Nuage de points des valeurs estimées en fonction des valeurs observées.

Construisons maintenant le graphe des résidus studentisés après la regression pas à pas et la sélection de nos 5 variables explicatives X1, X2, X4, X5 et X9 (le modèle simplifié).

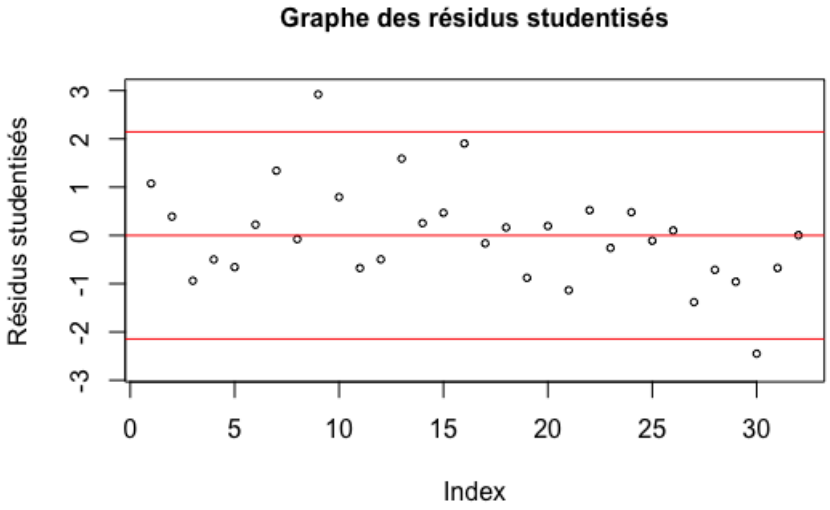

FIGURE 7 – Graphe des résidus studentisés après regressions pas à pas (seuils à 95%)

Comme pour celui obtenu avec les 9 variables explicatives, l'hypothèse d'homoscédacité est vérifiée, les résidus sont dispersés et correctement répartis autour de 0. Les jeux de données 9 et 30 sont encore en dehors de la bande de confiance à 95%.
Enfin, les tests de Shapiro-Wilk et de Durbin-Watson nous confirme la normalité ainsi que la non corrélation des résidus.

Pour finir ce TP nous avons essayé la commande *step* appliquée au modèle initial. Cette commande permet d'éliminer les variables explicatives pas à pas suivant le critère

d'information Akaike ou AIC. Ici on garde les variables X1, X2, X4, X5 et X9, ce que nous avions trouvé en effectuant les deux méthodes de régression manuelles plus tôt.
Ainsi, nous obtenons l'équation suivante pour modéliser nos données :

$$Y = 6.455 - 0.00255X1 - 0.0412X2 - 0.572X4 + 0.135X5 - 0.319X9$$

3 Conclusion

Pour conclure, les variables explicatives X1 (altitude en mètre), X2 (pente en degrés), X4 (hauteur en mètre des arbres), X5 (diamètre) et X9 (nombre de strates) permettent d'expliquer 66% des nombres de nids de chenilles par arbre. On peut même se restreindre à 3 ou 4 variables explicatives selon la précision que l'on veut obtenir. L'utilisation d'un nombre de variables explicatives plus important ne permet pas un gain suffisant par rapport à la contrainte que cela signifie.
Nous avons obtenu l'équation suivante, qui permet de modéliser Y, le nombre moyen de nids par arbre :

$$Y = 6.455 - 0.00255X1 - 0.0412X2 - 0.572X4 + 0.135X5 - 0.319X9$$

Annexe : code R

```
Nomfile = "http ://lmi2.insa-rouen.fr/ bportier/Data/chenille.txt"
chenille ¡-read.table(file=Nomfile, header=T,dec=".")
attach(chenille)
plot(chenille)
pairs(chenille,gap=0.05,cex=0.8,col="purple")
cor(chenille)

Oblige de passer par lm pour calculer le VIF mais on introduit pas encore le modele
reslm ¡- lm(Y X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 ,chenille)
library(car)
vif(reslm)

Coefficients du modele
model¡-lm(Y X1+ X2 + X3 + X4 + X5 + X7 + X8 + X9 + X10 ,chenille)
model$coefficients

summary(lm(Y X1 + X2 + X3 + X4 + X5 + X7 + X8 + X9 + X10))

TitreRes = "Graphe des résidus"
plot(residuals(model), xlab="Index", ylab="Résidus",main= TitreRes , cex=0.6, cex.lab=1,
cex.main = 1,cex.axis=1)
abline(0,0, col=2, lwd=2)

Titre1 = "Graphe des résidus studentisés"
plot(rstudent(model),xlab="Index", ylab="Résidus studentisés",main= Titre1 , cex=0.6,
cex.lab=1, cex.main = 1,cex.axis=1, ylim=c(-2.8,3))
p=9
seuil ¡- qt(0.975,n-p-2)
abline(h=c(-seuil , 0, seuil),col=2)

Titre = "Nuage de points (Observés/Prévus)"
plot(Y,predict(model), xlab="Nombre moyen de nid observé", ylab="Nombre moyen de
nid estimé",main= Titre , cex=0.8, cex.lab=1, cex.main = 1,cex.axis=1, ylim=c(-0.8,2))
abline(0,1, col=2, lwd=2)


shapiro.test(model$residuals)
hist(residuals(model),col="CC00CC",xlab="Résidus",ylab="Fréquences",main="Histogramme
des résidus",tck=0.01, freq=FALSE)
durbinWatsonTest(reslm)

Backward regression
```

```
summary(lm(Y X1+ X2 + X3 + X4 + X5 + X7 + X8 + X9 + X10))
summary(lm(Y X1+ X2 + X3 + X4 + X5 + X8 + X9 + X10))
summary(lm(Y X1+ X2 + X3 + X4 + X5 + X9 + X10))
```

```
summary(lm(Y X1+ X2 + X4 + X5 + X9 + X10))
summary(lm(Y X1+ X2 + X4 + X5 + X9 ))
summary(lm(Y X1+ X2 + X4 + X5 ))
summary(lm(Y X1+ X2 + X4 + X5))
```

Coefficients du nouveau modele

```
model¡-lm(Y X1+ X2 + X4 + X5 ,chenille)
model$coefficients
summary(lm(Y X1 + X2 + X4 + X5))
```

Forward regression

```
summary(lm(Y  X9 + X1 + X2 + X3))
summary(lm(Y  X9 + X1 + X2 + X4))
summary(lm(Y  X9 + X1 + X2 + X5))
summary(lm(Y  X9 + X1 + X2 + X7))
summary(lm(Y  X9 + X1 + X2 + X8))
```

Modèle simplifié

```
modelfinal¡-lm(Y X1 + X2 + X4 + X5 + X9 ,chenille)
modelfinal$coefficients
summary(lm(Y X1 + X2 + X4 + X5 + X9))
```

```
Titre = "Nuage de points (Observés/Prévus)"
plot(Y,predict(modelfinal), xlab="Nombre moyen de nid observé", ylab="Nombre moyen
de nid estimé",main= Titre , cex=0.8, cex.lab=1, cex.main = 1,cex.axis=1, ylim=c(-
0.8,2))
abline(0,1, col=2, lwd=2)
```

```
TitreRes = "Graphe des résidus"
plot(residuals(modelfinal), xlab="Index", ylab="Résidus",main= TitreRes , cex=1.5, cex.lab=1.6,
cex.main = 1.7,cex.axis=1.5)
abline(0,0, col=2, lwd=2)
```

```
Titre2 = "Graphe des résidus studentisés"
plot(rstudent(model),xlab="Index", ylab="Résidus studentisés",main= Titre2 , cex=0.6,
cex.lab=1, cex.main = 1,cex.axis=1, ylim=c(-2.8,3))
p=5
seuil ¡- qt(0.975,n-p-2)
abline(h=c(-seuil , 0, seuil),col=2)
```

```
shapiro.test(modelfinal$residuals)
durbinWatsonTest(reslm)
```

```
model¡-lm(Y X1+ X2 + X3 + X4 + X5 + X7 + X8 + X9 + X10 ,chenille)
step(model)
```