INSA – Génie Mathématique Année 2018-2019

GM5 – Régression Non Linéaire avec R – Feuille TP 1

Régression Linéaire Multiple

L'objectif de ce TP est de manipuler les instructions de base permettant de faire une régression linéaire simple et multiple et d'expérimenter les méthodes de sélection de variables et/ou de modèles.

Partie 1 Prise en main de la fonction `lm`

L'objet de cette première partie est de manipuler les instructions de base associées à la fonction `lm` en étudiant les données suivantes.

Un centre de tri reçoit chaque jour une importante quantité de sacs de courrier contenant des lettres et des paquets. Pour organiser l'activité du jour, on utilise le poids du courrier reçu la veille pour prédire le volume de courrier à trier, permettant ainsi de rationaliser le fonctionnement du service. Une liaison de type linéaire entre le poids du courrier reçu un jour donné et le nombre de lettres semble plausible. Afin de le justifier, on effectue une étude menée sur 21 jours choisis dans l'année en fonction du poids de courrier reçu et on mesure le nombre de lettres. correspondant. Ces données (ordonnées suivant le poids) sont stockées dans le fichier `courrier.txt` que l'on trouvera à l'URL
`http://lmi2.insa-rouen.fr/~bportier/Data/courrier.txt`

Pour les questions se reporter au fichier **TP-courrier.pdf** qui se trouve à l'URL :
`http://lmi2.insa-rouen.fr/~bportier/TP_courrier.pdf`

Partie 2 Régression linéaire multiple – Sélection de variables

L'objet de cette partie est d'expérimenter les méthodes de sélection de variables ascendante et descendante.

On considère les données provenant d'une étude réalisée en 1973 sur l'influence de certaines variables sur la densité de peuplement d'un parasite : la chenille processionnaire du pin (traité dans l'ouvrage de Tomassone et al., 1992).

La processionnaire du pin est un papillon nocturne de la famille des Notodontidés. La chenille se développe de préférence sur des pins et peut causer des dégâts considérables. On souhaite connaître l'influence de certaines caractéristiques de peuplements forestiers sur leurs développements.
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 :

- la variable à expliquer : Y : nombre moyen de nids par arbre
- les variables explicatives :
  - 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 échantillonnédans la placette
— X5 : le diamètre de cet arbre
— X6 : la note de densitéde peuplement
— X7 : l'orientation de la placette (de 1=sud à2=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)

Les données sont disponibles à l'URL :

`http://lmi2.insa-rouen.fr/~bportier/Data/chenille.txt`

1. Télécharger ces données à l'aide de la commande

   ```
   Nomfile =  "http://lmi2.insa-rouen.fr/~bportier/Data/chenille.txt"
   chenille <-read.table(file=Nomfile, header=T,dec=".")
   ```

   et taper la commande `attach(chenille)` pour définir les variables `Y, X1, X2, ...`

2. Y-a-t'il des corrélations fortes évidentes à l'oeil nu avec `Y` et entre les différentes variables explicatives ? On pourra utiliser la commande `plot(chenille)`. Comparer avec `pairs(chenille,gap=0.2,col="purple")`.
   Regarder ce que donne aussi `round(cor(chenille),2)`.

3. Etudier les multi-colinéarités à l'aide de la suite de commandes :

   ```
   > reslm <- lm(Y~X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 ,chenille)
   > library(car)
   > vif(reslm)
   ```

   Attention, on notera que la méthode de calcul nécessite l'utilisation de la fonction `lm`, mains en aucun cas, nous ne sommes en train d'utiliser une modélisation linéaire des données.

4. On envisage une liaison linéaire pour expliquer `Y` à partir des différentes variables explicatives. Quelles variables explicatives préconisez-vous de prendre ? Calculer les coefficients estimés du modèle à l'aide de la commande

   ```
   > model<-lm(Y~X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 ,chenille)
   > model$coefficients
   ```

   Bien évidemment, vous ne mettrez dans le modèle que les variables que vous aurez sélectionnées à l'issue de l'étape de statistique descriptive.
   Ces deux instructions sont équivalentes à

   ```
   > lm.chenille<-lm(Y~X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)
   > lm.chenillee[[1]]
   ```

   Ce qui change, c'est l'objet que l'on manipule après, soit `model`, soit `lm.chenille`.

5. Etudier et commenter la pertinence et la qualité de l'ajustement linéaire de ces données. Pour cela, on pourra
   — tester l'hypothèse de non régression ;

— calculer la part de variance expliquée par le modèle ;
— tracer le graphe des résidus ;
— étudier les résidus studentisés ;
— tracer le nuage de points "(observés, estimés)" ;
— tester la normalité des résidus à l'aide du test de Shapiro-Wilk. Pour cela, on utilisera l'instruction `shapiro.test(model$residuals)`.
— tester la non-corrélation des résidus à l'aide du test de Durbin-Watson. Pour cela, on utilisera l'instruction `durbinWatsonTest(reslm)` (fonction du package `car`).

6. On veut maintenant ne garder dans le modèle que les régresseurs significatifs, c'est à dire ceux qui ont une influence réelle sur la variable à expliquer `Y`. Déterminer la liste de ces variables en utilisant la méthode *forward regression*, puis la méthode *backward regression* :

   (a) Pour la régression pas à pas ascendante (forward regression), on commence par lancer toutes les régressions linéaires à 1 variable, en exécutant,
   - pour la variable `X1` la commande :
     `> summary(lm(Y ~X1))[[4]]`
   - pour la variable `X2` la commande :
     `> summary(lm(Y ~X2))[[4]]`

   — ...

   puis on sélectionne parmi les variables explicatives qui ont une influence réelle sur `Y`, celle qui est la plus significative, autrement dit celle qui a la p-value la plus petite et < 0.05. Puis, on effectue toutes les régressions linéaires à 2 variables explicatives comprenant la variable explicative qui vient d'être sélectionnée. On arrête le processus de sélection lorsqu'on ne trouve plus de régresseurs ayant une influence significative.

   (b) Pour la régression pas à pas descendante (backward regression), on commence par lancer la régression linéaire avec tous les régresseurs, en exécutant,

   ```
   > summary(lm(Y~X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10))[[4]]
   ```

   Puis, on sélectionne les variables qui n'ont pas une influence réelle sur `lpsa` et on élimine du modèle celle qui a la p-value la plus grande et ≥ 0.05. On relance alors la régression linéaire avec le modèle amputé de la variable éliminée et on réapplique le schema de sélection. On arrête le processus de sélection lorsqu'il n'y a plus de variables non significatives.

   (c) Quelle est la part de variance expliquée avec le modèle simplifié ? Comparer avec le modèle complet. On notera `modfinal` le modèle simplifié.

   (d) Etudier la qualité de l'ajustement en étudiant notamment les résidus, en recherchant d'éventuelles valeurs aberrantes, ...

   On pourra utiliser les commandes suivantes :

   ```
   plot(rstudent(modfinal))
   seuil <- qt(0.975,n-p-2)
   abline(h=c(-seuil , 0, seuil))
   lines(loewess(rstudent(modfinal)))
   plot(cooks.distance(modfinal),type="h")
   ```

7. Essayer les commandes

```
> model <- lm(Y~X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)
> step(model)
```

Commenter.

Indications pour la conception du rapport de TP : Deux masques sont à votre disposition aux adresses URL suivantes :

- `http://lmi2.insa-rouen.fr/~bportier/NomTP1.doc`
- `http://lmi2.insa-rouen.fr/~bportier/NomTP1.tex`

Vous êtes invités à les utiliser. Le corps principal du rapport ne devra pas excéder 10 pages. Il pourra être complété par une annexe contenant par exemple le code `R` qui a été élaboré ou toute information secondaire que vous jugerez pertinente de mettre dans le rapport. Pour me permettre une meilleure gestion de vos compte-rendus de TP, le nom de votre compte-rendu devra toujours être de la forme `NomTP*.pdf` où `Nom` désigne votre nom de famille et $*$ le numéro du TP réalisé.