TP Régression linéaire multiple et Modèle Linéaire Généralisé

Mathilde Boissel

25/01/2021

Table of Contents

Régression linéaire multiple

Exercice 1 : lecture des sorties

Nous allons utiliser le jeu de données `trees`, disponible dans le package `datasets` (nativement chargé dans R).

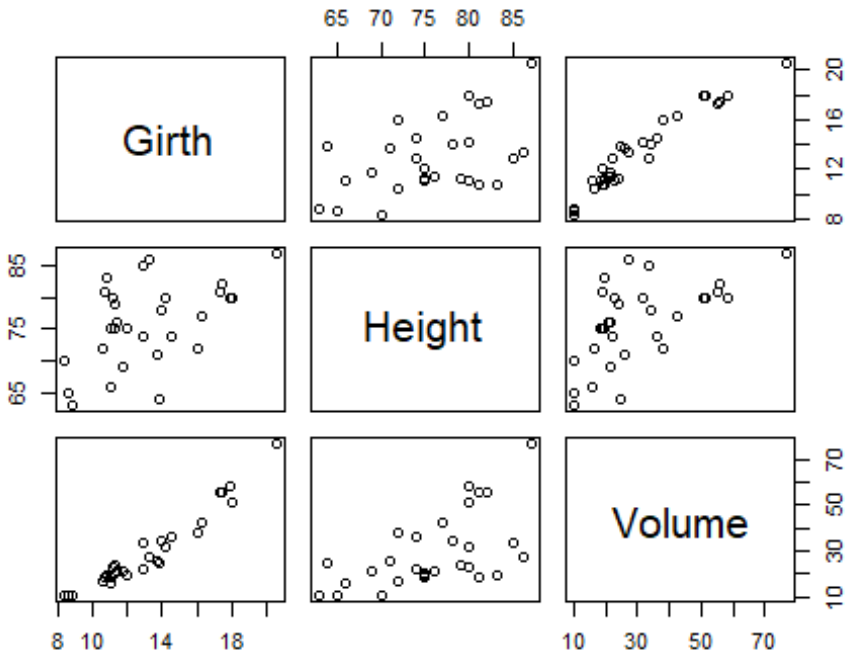
On souhaite expliquer la variable quantitative Volume (Volume of timber in cubic ft) à partir de 2 autres variables quantitatives Girth (Tree diameter (rather than girth, actually) in inches) et Height (Height in ft).

Pour se faire on considère le modèle rlm suivant :

$$Volume = \beta_0 + \beta_1 \times Girth + \beta_2 \times Height + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, les paramètres $\beta_0, \beta_1, \beta_2$ et σ sont des réels inconnus. On les estime alors avec n observations de $(Volume, Girth, Height)$ par la methède des mco. Un résumé et une visualisation des données est proposé ci-dessous :

```
# head(datasets::trees)
pairs(trees)
```



```r
str(trees)
```

```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

Puis le tableau de ce modèle rlm renvoyé par la commande `summary` est affiché ci-dessous :

```r
reg <- lm(formula = Volume ~ Girth + Height, data = trees)
summary(reg)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

1/ Quelle est la valeur de n ?

2/ Donner l'emco ponctuel de β_2.

3/ Est ce que la régression est hautement significative pour `Girth` ?

4/ Donner un intervalle de confiance pour β_2 à 95%.

5/ Donner le R2 et le R2 ajusté.

6/ Donner f_{obs} et la p-valeur du test global de Fisher. Quelle est l'hypothèse nulle associée à ce test statistique ?

7-A/ Donner la prédiction de Volume pour Girth valant 8.3 et Height valant 70 (avec la commande R et calculé "manuellement")

7-B/ Donner un intervalle de confiance à cette prédiction.

8/ Représenter le(s) graphique(s) des résidus. Est-ce que les hypothèses standards semblent être satisfaites ?

9/ Étude de la multicolinéarité : **Règle de Klein**

Pour chaque variable d'un rlm, 2 à 2, Si une ou plusieurs corrélations au carré sont proches du R2 du modèle, alors on soupçonne que les variables associées sont colinéaires.

Calculer le carré du coefficient de corrélation entre Girth et Height. Ces variables sont-elles colinéaires ?

10/ Y-a-t'il des valeurs aberrantes/extrêmes dans notre jeu de données ?

10-A/ Afin de détecter la présence de valeurs aberrantes, on peut utiliser une mesure nommée **Distance de Cook**. On envisage l'anormalité de la i-éme observation si $d_i > 1$.

Mais attention retirer strictement des valeurs sur ce seul critère serait une décision un peu rapide. Même "extrêmes", si les valeurs sont "vraiment" observées (mais ne nous arrangent pas), nous ne pouvons pas gommer un point pour améliorer le modèle.

Calculer les distances de cooks.

```
cook <- cooks.distance(reg)
cook[cook>1]
```

10-B/ Observations influentes : Pour identifier les observations qui influent le plus dans les estimations (celles dont une faible variation des valeurs induit une modification importante des estimations), plusieurs outils complémentaires existent :

les **DFBETAS** (bfb.[nom_de_variable]), les **DFFITS**, les **rapports de covariance** et les **distances de Cook**.

Si besoin est, pour identifier les observations influentes, on fait:

```
summary(influence.measures(reg))
```

Répondre à la question initiale.

Exercice 2 : Comparaison de 2 modèles

A notre jeu de données `trees`, nous ajoutons 2 nouvelles variables créées de toute pièce comme suit :

```
mydata <- trees
set.seed(25012021)
mydata$X3 <- rnorm(n = nrow(trees), mean = 30, sd = 1)
set.seed(25012021)
mydata$X4 <- rnorm(n = nrow(trees), mean = 60, sd = 3)
str(mydata)
```

```
## 'data.frame':    31 obs. of  5 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
##  $ X3    : num  28.2 29.5 30.5 30 30.7 ...
##  $ X4    : num  54.6 58.4 61.5 60 62 ...
```

N.B. : la fonction `set.seed()` vous permettra de rendre vos simulations reproductibles. Si vous utilisez des fonctions qui génèrent des nombres aléatoires (comme `rnorm()` ici), et que vous souhaitez partager les mêmes données avec d'autres personnes ou simplement retrouver les mêmes résultats une prochaine fois, il est important d'utiliser une graine = "seed" donnée.

Pour tester l'influence d'une ou plusieurs variables dans un modèle, tout en prenant en considération les autres variables, on peut utiliser le **test ANOVA** : si p-valeur > 0.05, alors les variables étudiées ne contribuent pas significativement au modèle.

Ici, on veut tester H0 : $\beta_3 = \beta_4 = 0$ en sachant qu'il y a toujours les variables Girth et Height dans le modèle. On effectue :

```
reg1 = lm(Volume ~ Girth + Height + X3 + X4, data = mydata)
reg2 = lm(Volume ~ Girth + Height, data = mydata)
anova(reg1, reg2)
```

```
## Analysis of Variance Table
##
## Model 1: Volume ~ Girth + Height + X3 + X4
## Model 2: Volume ~ Girth + Height
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 421.28
## 2     28 421.92 -1  -0.64611 0.0414 0.8403
```

1/ A la lecture de ces résultats, que pouvez-vous conclure ?

2/ Critères **AIC** et **BIC**

Ces critères AIC et BIC reposent sur un compromis "biais - parcimonie". Plus petits ils sont, meilleur est le modèle.

```r
message("reg1")
message("AIC = ", AIC(reg1))
## AIC = 176.056853855826
message("BIC = ", BIC(reg1))
## BIC = 184.660777082737
message("reg2")
message("AIC = ", AIC(reg2))
## AIC = 176.90997298727
message("BIC = ", BIC(reg2))
## BIC = 182.645921805211
```

Votre conclusion change-t-elle avec ces nouveaux résultats ?

GLM : Régression logistique

On considère une population P divisée en 2 groupes d'individus G1 et G2 distinguables par des variables $X_1, \ldots, X_p$. Soit Y la variable qualitative valant 1 si l'individu considéré appartient à G1 et 0 sinon. On souhaite expliquer Y à partir de $X_1, \ldots, X_p$.

Dans le cadre d'un Régression logistique, on souhaite estimer la probabilité qu'un individu i vérifiant $(X_1, \ldots, X_p) = x$ appartienne au groupe G1 :

$$p(x) = \mathbb{P}(\{Y = 1\}|x) = \mathbb{E}(Y|x)$$

$$p(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

La transformation logit s'applique donc dans ce cas.

Exercice 3 : Régression logistique simple

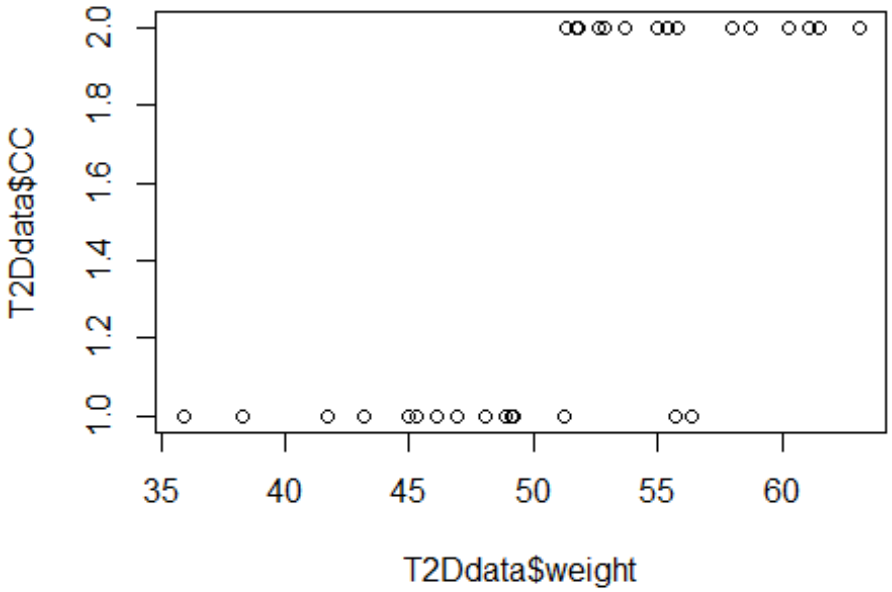
Dans cette configuration, nous utiliserons des données `T2Ddata`, simulées comme suit :

- `CC` est la variable binaire qui traduit le statut "cas" (1) pour définir le groupe des enfants diabétiques ou "ctrl", controle, (0) pour définir les enfants non diabétiques.
- `weight` la variable numérique qui représente le poids des individus.

```
T2Ddata <- data.frame(
  weight = c(35.9, 38.3, 55.7, 41.7, 43.2, 49.1, 45, 45.3, 46.1, 46.9, 48.1,
    48.9, 49.2, 51.2, 56.4, 51.7, 51.8, 52.6, 52.9, 51.3, 53.7, 55,
    55.4, 55.8, 58, 58.7, 60.3, 61.1, 61.5, 63.1),
  CC = factor(x = c(rep("0", 15), rep("1", 15)), levels = c("0", "1"), labels
= c("CTRL", "CAS"))
)
str(T2Ddata)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  35.9 38.3 55.7 41.7 43.2 49.1 45 45.3 46.1 46.9 ...
##  $ CC    : Factor w/ 2 levels "CTRL","CAS": 1 1 1 1 1 1 1 1 1 1 ...
```

1/ Visualiser les données



2/ Réaliser la régression logistique modélisant `CC` en fonction de `weight`.

3/ **Rapport des côtes** ou **odds ratio**

Si X_j augmente d'une unité, alors le rapport des côtes est $RC_j = \exp(\beta_j)$.

Par conséquent,

- si $RC_j > 1$, l'augmentation d'une unité de Xj entraîne une augmentation des chances que {Y = 1} se réalise,
- si $RC_j = 1$, l'augmentation d'une unité de Xj n'a pas d'impact sur Y,
- si $RC_j < 1$, l'augmentation d'une unité de Xj entraîne une augmentation des chances que {Y = 0} se réalise.

Calculer l'odd ration de weight.

4/ Avec ce nouveau jeu de données `T2Ddata2`, refaites les mêmes opération (question 1 à 3). Que se passe-t-il ?

```
T2Ddata2 <- data.frame(
  weight = sort(c(35.9, 38.3, 55.7, 41.7, 43.2, 49.1, 45, 45.3, 46.1, 46.9, 48.1,
    48.9, 49.2, 51.2, 56.4, 51.7, 51.8, 52.6, 52.9, 51.3, 53.7, 55, 55.4, 55.8, 58, 58.7, 60.3, 61.1, 61.5, 63.1)),
  CC = factor(x = c(rep("0", 15), rep("1", 15)), levels = c("0", "1"), labels = c("CTRL", "CAS"))
)
str(T2Ddata2)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  35.9 38.3 41.7 43.2 45 45.3 46.1 46.9 48.1 48.9 ...
##  $ CC    : Factor w/ 2 levels "CTRL","CAS": 1 1 1 1 1 1 1 1 1 1 ...
```

Exercice 4 : Régression logistique multiple

Nous allons utiliser le jeu de données `esoph`, disponible dans le package `datasets` (nativement chargé dans R).

Soit $j \in \{0, \ldots, p\}$. Le **test de la déviance** vise à évaluer l'influence (ou la contribution) de Xj sur Y. La p-valeur associée utilise la loi du Chi-deux : si *, l'influence de Xj sur Y est significative, si **, elle est très significative et si ***, elle est hautement significative.

Ici, on souhaite modéliser la proportion d'individus cas/control définit dans les 2 colonnes ncases et ncontrols.

Evaluer les 2 modèles suivants et noter les différences.

```r
str(esoph)

model1 <- glm(
  cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp,
  data = esoph, family = binomial(link = "logit")
)
summary(model1)
anova(model1, test = "Chisq")

model2 <- glm(
  cbind(ncases, ncontrols) ~ agegp + unclass(tobgp) + unclass(alcgp),
  data = esoph, family = binomial()
)
summary(model2)
anova(model2, test = "Chisq")
```

Pour aller plus loin

Complément

Les éléments ci-dessous ne seront pas forcement testés en TP (ni vu en cours). Mais vous pourrez commencer à vous familiariser avec ces concepts grâces à ces quelques remarques.
De plus si vous voulez aller plus loin, je vous recommande de réaliser les études proposé ici :

https://chesneau.users.lmno.cnrs.fr/etudes-reg.pdf

Le cours allant avec est aussi disponible ici :

https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf

Multicolinéarité

la variance de $\hat{\beta}_j$ explose et entraéne une grande instabilité dans l'estimation de β_j et fausse tous les tests statistiques.

En particulier, si au moins une variable parmi $X_1, \ldots, X_p$ a une liaison linéaire avec d'autres, il est possible qu'aucune variable ne montre d'influence significative sur Y et cela, en dépit de toute logique, et du test de Fisher qui peut quand même indiquer une influence significative globale des coefficients (car il prend en compte toutes les variables).

Règle de Klein

Si une ou plusieurs valeurs au carré sont proches de R2, alors on soupçonne que les variables associées sont colinéaires.

```
c = cor(cbind(X1, X2, X3), cbind(X1, X2, X3))
c^2
```

VIF

On appelle **vif** V_j le facteur d'inflation de la variance associé à la variable X_j. Si $V_j \geq 5$, alors on admet que X_j a un lien linéaire avec les autres variables.

```
library(car)
vif(reg)
```

Sélection de variables

Il est intéressant de déterminer la meilleure combinaison des variables X1, …, Xp qui explique Y. Or l'approche qui consiste à éliminer d'un seul coup les variables non significatives n'est pas bonne ; certaines variables peuvent être corrélées à d'autres, ce qui peut masquer leur réelle influence sur Y.

Plusieurs approches sont possibles :

- Approche exhaustive,
- Approche en arrière,
- Approche en avant,
- Approche pas à pas.

Rappel sur les critères Cp, AIC et BIC :

Ces critères Cp de Mallows, AIC et BIC reposent sur un compromis “biais - parcimonie”. Plus petits ils sont, meilleur est le modèle.

```r
AIC(reg)
BIC(reg)

library(leaps)
v = regsubsets(Y ~ X1 + X2 + X3, w, method = "backward")
plot(v, scale = "bic")
# Notons que l'option scale = "aic" n'existe pas. On obtiendrait toutefois la
méme sélection
# de variables que celle obtenue avec l'option scale = "bic"

library(stats)
# Pour utiliser l'approche pas à pas avec le AIC, puis obtenir les résultats
statistiques associés au modèle sélectionné :
reg2 = step(reg, direction = "both", k = 2)
summary(reg2)
# Pour considérer le BIC, on prend k = log(length(Y))
```