````
---
title: "TP Régression linéaire multiple et Modèle Linéaire Généralisé"
author: "Mathilde Boissel"
date: "25/01/2021"
output: word_document
toc: true
toc_depth: 2
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
```

```{r logo-udl, results="asis", echo = FALSE, fig.align='center'}
knitr::include_graphics("./Images_includ/logo_univ-lille-large.png")
```

\newpage

# Régression Linéaire Multiple (rlm)

## 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 $\sigma$ sont des réels inconnus. On les estime alors avec $n$ observations de $(Volume, Girth, Height)$ par la méthode des mco (moindres carrées ordinaires).

=> Exécuter les commandes et répondre aux questions.

Un résumé et une visualisation des données est proposé ci-dessous :

```{r exo1, eval = TRUE, echo = TRUE}
# head(datasets::trees)
pairs(trees)
str(trees)
```

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

```{r exo1_reg, eval = TRUE, echo = TRUE}
reg <- lm(formula = Volume ~ Girth + Height, data = trees)
summary(reg)
```

1/ Quelle est la valeur de $n$ ?
<!-- n=31, info vu dans le str, ou déductible avec les d.d.l. -->
<!-- `nrow(trees)` donne aussi le nombre de ligne dans un dataset -->

2/ Donner l'emco ponctuel de $\beta_2$.
<!-- beta_2 concerne Height = 0.3393 -->
<!-- `coef(reg)` retourne les coefficents, coef(reg)[3] donnera directement Beta2 -->
<!-- ou  `res <- summary(reg)` et faire `res$coefficients` -->

3/ Est-ce que la régression est hautement significative pour `Girth` ?
<!-- A lire dans le sortie du summary(reg) -->
<!-- hautement significatif oui car p-valeur <0.001, symbolisé par *** -->

4/ Donner un intervalle de confiance pour $\beta_2$ à 95%.
````

```{r confint, echo = FALSE, eval = FALSE}
confint(reg, level = 0.95)
# ou plus précisement uniquement :
confint(reg, level = 0.95)[3, ]
```

5/ Donner le R2 et le R2 ajusté.
<!-- a lire avec `summary(reg)` -->
<!-- Multiple R-squared: 0.948,   Adjusted R-squared: 0.9442 -->
<!-- On peut aussi les obtenir comme suit : -->
<!-- res <- summary(reg) -->
<!-- `res$r.squared` et `res$adj.r.squared` -->

6/ Donner f_{obs} et la p-valeur du test global de Fisher. Quelle est l'hypothèse nulle associée à ce test statistique ?
<!-- a lire avec `summary(reg)` -->
<!-- F-statistic: 255 et sa p-value: < 2.2e-16 -->
<!-- H0 : tous les beta = $\mathbf{0}$ contre H1 : "il existe au moins un coefficient non nul". -->
<!-- ici on rejette H0, cela signifie que le modele est pertinent -->

<!-- La statistique de Fisher peut aussi se relire ici : `res$fstatistic`, -->
<!-- Avec la fonction `pf`, on pourrait aller retrouver la p-valeur exact de Fisher : `pf(q = res$fstatistic[["value"]],df1 = res$fstatistic[["numdf"]], df2 = res$fstatistic[["dendf"]], lower.tail=FALSE)` -->

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

```{r predictA, echo = FALSE, eval = FALSE}
## avec R :
predict(reg, data.frame(Girth = 8.3, Height = 70))
# predict(reg)[1] ## On remarque que c'est justement la première observation du jeu de données, et predict se base par defaut sur le jeu de données utilisé pour la Régression.

## Manuellement : prediction de y_x = b0 * 1 + b1 * Girth + b2 * Height
sum(reg$coefficients * c(const = 1, Girth = 8.3, Height = 70))
```

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

```{r predictB, echo = FALSE, eval = FALSE}
predict(reg, data.frame(Girth = 8.3, Height = 70), interval = "confidence")
```

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

```{r residus, echo = FALSE, eval = FALSE}
e = residuals(reg)
plot(e)
abline(h = 0, col = "red")
## ou directement :
par(mfrow = c(2, 2))
plot(reg)
## Pour faire un zoom sur le premier graph qui semble avoir des probèmes :
par(mfrow = c(1,1)); plot(reg, 1)
## une structure est bien visible, nous n'avons pas une symétrie autour de l'axe y = 0. On pourrait supposer que les résidus sont dépendants
# par(mfrow = c(1,1)); plot(reg, 2)

## pour aller plus loin sur l'indendances des erreurs :
par(mfrow = c(1, 2))
acf(e)
pacf(e)

par(mfrow = c(1, 1)) ## pour remettre les params graphic
```

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

Pour chaque variable d'un modèle de 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.
Soupçonne-t-on ces variables d'être colinéaires ?

```{r multicor, echo = FALSE, eval = FALSE}
cor(x = trees$Girth, y = trees$Height)^2
# A comparer avec : summary(reg)$r.squared
```

<!-- Cela renvoie 0.26, lequel est éloigné de R2 = 0.948. -->
<!-- Donc, par la règle de Klein, il n'y a pas de lien linéaire entre Girth et Height. -->

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.

```{r cook, echo = TRUE, eval = FALSE}
cook <- cooks.distance(reg)
cook[cook>1]
```

<!-- aucune distance > 1, pourtant le graphique des résidus `plot(e)`(plus haut) montrait un point un peu éloigné des autres... (le 31). On peut donc regarder d'autres critères (10-B) -->

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 :

```{r influence, echo = TRUE, eval = FALSE}
summary(influence.measures(reg))
```

Répondre à la question initiale.

<!-- Réponse en demi-teinte car avec le seul critère des distances de cook, aucune valeur n'était abérrante. Mais on voit tout de même que l'obersation 31 est "influente" (on dit alors qu'elle peut "driver" le signal, i.e. diriger (ou tirer vers le haut ou la bas) nos coefficients - la pente de la droite de regression) -->
<!-- Il serait bon de regarder plus attentivement si l'observation "31" se distingue "trop" des autres dans les graphiques des données (type "pairs") et peut etre la retirer de notre modélisation -->

\newpage

Exercice 2 : Comparaison de 2 modèles

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

```{r exo2, eval = TRUE, echo = TRUE}
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)
```

N.B. : la fonction `set.seed()` vous permettra de rendre vos simulations reproductibles. Si vous utilisez des fonctions qui généèent 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" à choisir.

Pour tester l'influence d'une ou plusieurs variables dans un modèle alternatif, tout en prenant en considération les autres variables, on peut utiliser le **test ANOVA** : Au seuil 5% e.g.
Si p-valeur > 0.05, alors les variables étudiées dans le modèle alternatif ne contribuent pas significativement à expliquer notre variable d'intérêt, comparativement au modèle de base".

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 :

```{r anova, echo = TRUE, eval = TRUE}
reg1 = lm(Volume ~ Girth + Height + X3 + X4, data = mydata)
reg2 = lm(Volume ~ Girth + Height, data = mydata)
anova(reg1, reg2)
```

Reg1 est le modèle alternatif testé, Reg2 est le modèle de base.
1/ A la lecture de ces résultats, que pouvez-vous conclure ?

<!-- On lit la p-valeur = 0.8403 -->
<!-- si p-valeur > 0.05, alors les variables étudiées ne contribuent pas significativement au modèle. -->
<!-- ici le meilleur modèle est le modele 2 : X3 et X4 ne contribuent pas significativement au modèle... (comme attendu!)-->

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 AICBIC, echo = TRUE, eval=TRUE}
message("reg1")
message("AIC = ", AIC(reg1))
message("BIC = ", BIC(reg1))

message("reg2")
message("AIC = ", AIC(reg2))
message("BIC = ", BIC(reg2))
```

Votre conclusion change-t-elle avec ces nouveaux résultats ?
<!-- Non, le modèle Reg2 a toujours les plus petites valeurs. -->

\newpage

GLM : Cas Binomial - Régression Logistique

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

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

$$p(x) = \mathbb{P}(\{Y=1\}|x) = \mathbb{E}(Y|x)$$
$$p(x) = \beta_0 + \beta_1x_1 + ... + \beta_px_p$$
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 qualitative binaire qui traduit le statut "cas" (1) pour définir le groupe des individus diabétiques et "ctrl", contrôle, (0) pour définir les individus non diabétiques.

+ `weight` la variable numérique qui représente le poids des individus.

```{r T2Ddata, echo = TRUE, eval = TRUE}
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)
```

1/ Visualiser les données

```{r viz, echo = FALSE, eval = TRUE}
plot(T2Ddata$weight, T2Ddata$CC)
```

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

```{r glmlog, eval = FALSE, echo = FALSE}
reg <- glm(formula = CC ~ weight, data = T2Ddata, family = binomial(link = "logit"))
summary(reg)
```

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

**Définition : i X_j augmente d'une unité, alors le rapport des côtes est $RC_j = \mathbb{exp}
(\beta_j)$.**

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**,
En effet, si $\beta_j = 0$, alors $RC_j = \mathbb{exp}(0) = 1$
+ si **RC_j < 1**, l'augmentation d'une unité de Xj entraîne une **augmentation des chances que
{Y = 0} se réalise**.

=> Calculer l'odd ratio de weight.

<!-- Dans R, on obtient ces valeurs comme suit : -->

```{r OR, eval = FALSE, echo = FALSE}
OR = exp(coef(reg))
OR>1
```

<!-- L'OR de weight est > 1 donc l'augmentation d'une unité de weight entraîne une augmentation des
chances que {CC = 1} se réalise, i.e. l'augmentation du poids augmente le risque de venir
diabetique, selon nos données. -->

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

```{r T2Ddata2, echo = TRUE, eval = TRUE}
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)
```

<!-- Executer les commandes suivantes -->

```{r glmlog2, eval = FALSE, echo = FALSE}
plot(T2Ddata2$weight, T2Ddata2$CC)
reg2 <- glm(formula = CC ~ weight, data = T2Ddata2, family = binomial(link = "logit"))
summary(reg2)
```

```
OR2 = exp(coef(reg2))
OR2
```

<!-- Le graphique nous montre une très nette partition entre les mesures (on pourrait tracer un trait net vertical) -->
<!-- A la réalisation du GLM : Le warning suivant s'affiche : -->
<!-- Warning messages: -->
<!-- 1: glm.fit: l'algorithme n'a pas convergé -->
<!-- 2: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1 -->
<!-- Vu les données (construites exprès) ce warning était attendu. Mais il peut arriver aussi sur des données réelles, il faudra alors en garder la trace et ne pas seulement retourner les estimations -->
<!-- L'OR de weight "explose" (evidement puisqu'il n'a pas de sens, au vu du warning) -->

<!-- Ici la modélisation n'était pas pertinente car la partition entre les 2 groupes était totale selon la mesure utilisée dans le modèle. On s'en rend un peu mieux compte avec le boxplot suivant : `boxplot(T2Ddata2$weight~T2Ddata2$CC)` -->

\newpage

Exercice 4 : Régression logistique multiple

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

Soit $j \in \{0,..., p\}$. 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 esoph, echo = TRUE, eval = FALSE}
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")
```

<!-- Modele 1 effets avec interaction de alcohol et tobacco (tobgp * alcgp) + groupe d'age -->
<!-- Modele 2 test un effet linéaire de alcohol et tobacco (avec unclass()) et sans interaction. -->
<!-- Ceci est fait pour vous montrer qu'on peut parfois se demander si une interaction est necessaire... on pourra ici le tester. et aussi pour vous montrer que la classe des variables est importante et joue un role dans leur considération dans le modèle (dans les estimations de vos coefs) -->

\newpage

Pour aller plus loin

Compleéent

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 $X_1, ..., X_p$ 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).

Réèle de Klein

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

```{r, klein, echo = TRUE, eval = FALSE}
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 $V_j \ge 5$, alors on admet que X_j a un lien linéaire avec les autres variables.

```{r, vif, echo = TRUE, eval = FALSE}
library(car)
vif(reg)
```

Sélection de variables

Il est intéressant de déterminer la meilleure combinaison des variables X1, ..., Xp qui explique Y. r l’approche qui consiste à éliminer d’un seul coup les variables non significatives n’est pas bonne ; ertaines 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 selection, echo = TRUE, eval = FALSE}
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))
```