

Travail Pratique 4

La régression linéaire simple avec R

STT1039
BIOLOGIE QUANTITATIVE

Réalisé par
Marc Pépino



Département des sciences de l'environnement

1. Objectifs du TP4

Le travail pratique 4 (TP 4) a pour objectif général d'initier les étudiants à la régression linéaire simple réalisée à l'aide du logiciel R. À la fin du TP 4, l'étudiant doit savoir : effectuer une régression linéaire simple entre deux variables quantitatives, vérifier les conditions d'application de la régression linéaire simple à l'aide de l'examen des résidus, illustrer la relation entre les deux variables à l'aide de la droite de régression, extraire les coefficients, la valeur de p et le coefficient de détermination.

Note : Tout ce qui est mis en rouge et en gras dans les exemples (encadrés) sont des noms donnés par l'utilisateur. Ces noms sont à modifier en conséquence si les exemples sont utilisés avec d'autres bases de données.

Les tests statistiques présentés dans les exemples seront réalisés à partir de la base de données *TP4_Exemples.csv* disponible sur le portail du cours :

```
> #Importation de la base de données
> tab <- read.csv("D:/bureau/TP4_Exemples.csv",sep=";",dec=".",h=T)
>
> #Visualisation de la structure du tableau de données
> str(tab)
'data.frame': 60 obs. of 2 variables:
 $ x : num 190 188 208 200 206 ...
 $ y : num 633 552 677 657 694 ...
```

OPTIONNEL :

```
#Une base de données similaire peut être obtenue à partir du code suivant:
#Construction des variables
n <- 60
x <- rnorm(n,200,10)
y <- rep(NA,100)
y <- 55 + 3*x + rnorm(n,0,30)

#Construction du tableau de données
tab <- as.data.frame(x)
tab$y <- y
```

Pour plus de détails sur les tests statistiques avec R, vous pouvez consulter, entre autres, **Crawley (2015)**, **Millot (2018)**, ou **Bertrand et Maumy-Bertrand (2010)**.

2. Rappel : la corrélation de Pearson

La fonction **cor()** permet de calculer le coefficient de corrélation de Pearson entre deux variables quantitatives (voir TP3). La fonction **cor.test()** permet d'effectuer le test sur ce coefficient.

```
> #Exemple 1
> #corrélation de Pearson
> cor.test(x,y,method="pearson",data=tab)
    Pearson's product-moment correlation

data: x and y
t = 6.5845, df = 58, p-value = 1.456e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4797914  0.7786453
sample estimates:
      cor 
0.6540345
```

3. La régression linéaire simple

La fonction **lm()** permet de faire la régression linéaire simple entre deux variables quantitatives. La fonction **summary()** permet d'obtenir le tableau de sortie de l'analyse.

```
> #Exemple 2
> #régression linéaire simple
> mod <- lm(y~x,data=tab)
> summary(mod)
Call:
lm(formula = y ~ x, data = tab)
Residuals:
    Min     1Q  Median     3Q    Max
-75.459 -16.878  1.487  17.391  62.902

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.7338    98.1005   0.160   0.873
tab$x         3.2044     0.4866   6.585 1.46e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.95 on 58 degrees of freedom
Multiple R-squared:  0.4278,    Adjusted R-squared:  0.4179 
F-statistic: 43.36 on 1 and 58 DF, p-value: 1.456e-08
```

Interprétation de la sortie de l'analyse :

Coefficients donne les informations sur l'ordonnée à l'origine (première ligne) et la pente (deuxième ligne) de l'équation de régression. La colonne **Estimate** donne les coefficients de la droite de régression. La colonne **Std. Error** donne l'erreur type des coefficients. La dernière colonne **Pr(>|t|)** donne la valeur de p pour le test sur l'ordonnée à l'origine (première ligne) et sur la pente (deuxième ligne). **Multiple R-squared** représente le coefficient de détermination (R^2).

L'exemple 3 ci-dessous montre trois façons différentes d'effectuer la même analyse :

```
#Exemple 3
#Première façon d'écrire le modèle à partir du tableau de données
mod1 <- lm(y~x,data=tab)

#Deuxième façon d'écrire le modèle à partir du tableau de données
mod2 <- lm(tab$y~tab$x)

#troisième façon d'écrire le modèle à partir de vecteurs
x <- tab$x
y <- tab$y
mod3 <- lm(y~x)
```

4. Condition d'applications

4.1 La normalité des résidus

La régression linéaire simple suppose que les résidus se distribuent suivant une loi normale de moyenne nulle. La visualisation des résidus sous forme d'histogramme (voir TP2) permet de juger de la violation de cette condition. Il est toutefois plus facile de juger de cette condition à partir du diagramme quantile-quantile (« Normal QQ-plot »). Le test de normalité de Shapiro-Wilk peut être utilisé pour appuyer l'interprétation graphique. La fonction **residuals()** permet de récupérer les résidus du modèle de la régression linéaire simple.

```
#Exemple 4
#regression linéaire simple
mod <- lm(y~x,data=tab)
#histogramme des résidus
par(mfrow=c(1,2))
hist(residuals(mod))
qqnorm(residuals(mod))
qqline(residuals(mod))
shapiro.test(residuals(mod))
```

4.2 La linéarité de la fonction de régression et égalité des variances conditionnelles

La visualisation de la distribution des résidus suivant les valeurs prédites par le modèle de régression permet de juger de la linéarité de la fonction de régression. La distribution des résidus ne doit montrer aucune forme particulière. Ce graphique permet également de vérifier visuellement l'égalité des variances des distributions conditionnelles.

```
#Exemple 5
#regression linéaire simple
mod <- lm(y~x,data=tab)
#résidus en fonction de x
x11()
plot(tab$x,residuals(mod))
abline(h=0, lty=2)
#Optionnel: ajouter une courbe de lissage
lines(lowess(residuals(mod)~tab$x),lwd=2)
```

Information complémentaire :

Le logiciel R fournit également plusieurs graphiques diagnostics afin de valider si les conditions d'applications du modèle sont respectées. La fonction **plot()** appliquée au modèle de régression permet de générer ces graphiques. Voir la partie théorique du cours pour une interprétation de ces graphiques.

4.3 Données exceptionnelles ou aberrantes

Les données exceptionnelles ou aberrantes peuvent se voir sur le diagramme de dispersion de y en fonction de x. La distance de Cook est une mesure quantitative qui permet de voir l'influence d'un point sur la droite de régression (effet de levier). Sur ce type de graphique, on ne veut pas voir des valeurs très élevées par rapport aux autres. On peut aussi déterminer une valeur seuil acceptable en prenant le quantile 50% de l'aire sous la courbe de la loi de Fisher –Snedecor à 2 et n-2 d.d.l.

```
#Exemple 6
#regression linéaire simple
mod <- lm(y~x,data=tab)
#Distance de Cook
x11()
plot(mod,which=4)
qf(0.5,2,length(tab$x)-2) #Valeur seuil
```

5. Illustration graphique de la régression linéaire simple

La première représentation à faire est le graphique montrant les données brutes entre les deux variables quantitatives (y en fonction de x). La fonction **abline()** appliquée au modèle permet de représenter la droite de régression sur le graphique. La fonction **predict()** permet de calculer l'intervalle de confiance de y pour de nouvelles valeur de x et d'ajouter ensuite ces intervalles de confiance sur le graphique. L'exemple 6 ci-dessous illustre comment ajouter la droite de régression et l'intervalle de confiance à 95%. Consulter le TP2 pour plus de détails sur les options graphiques.

```
#Exemple 7
#Régression linéaire simple
model <- lm(y~x,data=tab)

#Représentation graphique
x11()
plot(tab$x,tab$y)

#Droite de régression
abline(model)

#Intervalle de confiance à 95%
model <- lm(y~x,data=tab)
xnew <- data.frame(x = seq(min(tab$x),max(tab$x),length=100))
pred <- predict(model,xnew,interval="confidence",level=0.95)
lines(xnew[,1],pred[,2],lty=2,col="red")
lines(xnew[,1],pred[,3],lty=2,col="red")
```

6. Les fonctions complémentaires de la régression linéaire simple

Le tableau suivant reprend les principales fonctions qui peuvent être utilisées sur un modèle de régression. Consulter l'aide en ligne pour plus d'information sur ces fonctions et leur utilisation.

Tableau 1. Fonctions associées au modèle de régression linéaire simple (Extrait de Bertrand et Maummy-Bertrand 2010).

Fonction	Définition
summary(mod)	Tableau de sortie de l'analyse avec les tests statistiques
coef(mod)	Coefficients de la régression
fitted(mod)	Valeurs prédites pour les valeurs observées de x
predict(mod)	Valeurs prédites et intervalles de confiance pour de nouvelles valeurs de x
confint(mod)	Intervalle de confiance pour l'ordonnée à l'origine et la pente de la régression
residuals(mod)	Résidus de la régression

7. Exercices

Répondez aux exercices dans un fichier texte « Nom_Prenom_TP4.txt ». Au début du fichier, identifiez-vous par votre nom et prénom, ajouter la date et le numéro du TP, le tout en utilisant des commentaires. Utilisez des commentaires pour structurer vos réponses.

Considérer les deux variables suivantes qui représentent les données de longueur et de poids :

```
#Variables Long et Poids  
Long <- rnorm(50,220,25)  
Poids <- 300 + 3 * Long + rnorm(50,0,150)
```

Exercice 1. La relation entre la longueur et le poids est-elle significative ? Quelle est la valeur des coefficients de la droite de régression. Quel est le coefficient de détermination ? Recalculer le coefficient de détermination à partir de sa formule (variation expliquée divisé par la variation totale).

Exercice 2. Vérifier la normalité des résidus et la linéarité de la fonction de régression. Y-a-t-il des données exceptionnelles ou aberrantes ?

Exercice 3. Illustrer la relation entre la longueur et le poids. Ajouter la droite de régression et l'intervalle de confiance à 95% de la moyenne.

Bibliographie

- Bertrand, F. et Maumy-Bertrand, M. 2010. Initiation à la statistique avec R. Dunod, Paris, France.
- Crawley, M. J. 2013. The R book (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Crawley, M. J. 2015. Statistics: an introduction using R (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Legendre, P. 2009. Introduction au langage statistique R.
http://biol09.biol.umontreal.ca/BIO6077/Introduction_R.pdf.
- Millot, G. 2018. Comprendre et réaliser les tests statistiques à l'aide de R : manuel pour les débutants (4^{ème} éd.). De Boeck, Bruxelles, Belgique.
- Murrell, P. 2011. R graphics (2^{ème} éd.). Chapman & Hall/CRC The R series, Boca Raton, Floride.
- Paradis, E. 2005. R pour les débutants.
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.
- Short, T. 2004. R Reference Card. <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>.

Liens utiles :

- Site officiel de R.
<http://www.r-project.org>
- Ateliers R du CSBQ.
<https://qcbs.ca/wiki/r>
- Quick R.
<http://www.statmethods.net/>
- Apprentissage en ligne du logiciel R.
<https://www.datacamp.com/>
- Beginners statistics introduction.
http://influentialpoints.com/Training/basic_statistics_introduction.htm
- R Graphics Cookbook.
<http://www.cookbook-r.com/Graphs/>
- Liste des bases de données disponibles sous R.
<https://vincentarelbundock.github.io/Rdatasets/datasets.html>