I. CONDITIONS D'APPLICATION DU MODELE LINEAIRE SIMPLE

Hypothèses fondamentales du modèle linéaire simple $Y_i = \beta_1 + \beta_0 X_i + \varepsilon_i$

Où Y est la variable dépendante et X la variable explicative

Le modèle linéaire simple s'appuie sur les hypothèses suivantes :

1. Il existe une **relation de corrélation** (mais pas forcément de causalité !) entre une variable aléatoire dépendante Y et une variable explicative X.

   Hypothèse de corrélation

2. Les valeurs Xi prises par la variable X sont rigoureusement exactes, c.à d. : Var(Xi) = 0, ∀i.

   Hypothèse d'exactitude

3. La courbe joignant les **moyennes** des distributions des Yi pour les valeurs des Xi est appelée courbe ou équation de régression

   Hypothèse de linéarité entre X et Y

   Lorsque ces moyennes sont alignées, la courbe est alors une droite. Dans ce cas, l'équation de régression est de la forme : $E(Y_i) = \beta_1 + \beta_0 X_i$. On utilise aussi la notation E(Yi|Xi) pour indiquer que ces moyennes sont conditionnelles.

4. La variance σ² de chaque distribution des Yi est la même, quelque soient les valeurs de Xi prises par la variable explicative X, c.à.d. Var(Yi) = σ² pour tout X. Ceci signifie que la variance des erreurs (ε_i) demeure constante pour tous les Xi :

   Variance constante

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \mathrm{Var}(\beta_0 + \beta_1 X_i) + \mathrm{Var}(\varepsilon_i)$$
$$= \quad 0 \quad + \quad \sigma^2 \qquad (1)$$

   On suppose donc que l'ampleur de la dispersion de chaque distribution des Yi est identique, peu importe les valeurs prises par la variable explicative X.

5. Les Yi ne sont pas corrélés entre eux, c.à.d. que les observations de la variable dépendante ne sont aucunement liées avec les précédentes et n'influent pas les suivantes

   Absence de corrélation entre les Yi

6. Pour chaque valeur de Xi, les valeurs de Y réparties autour de l'équation de régression : $E(Y_i) = \beta_1 + \beta_0 X_i$ sont distribuées selon la loi normale. De façon équivalente, les ε_i sont distribuée normalement (et de même variance : cf(1)), de moyenne nulle.

   Distribution normale des Yi

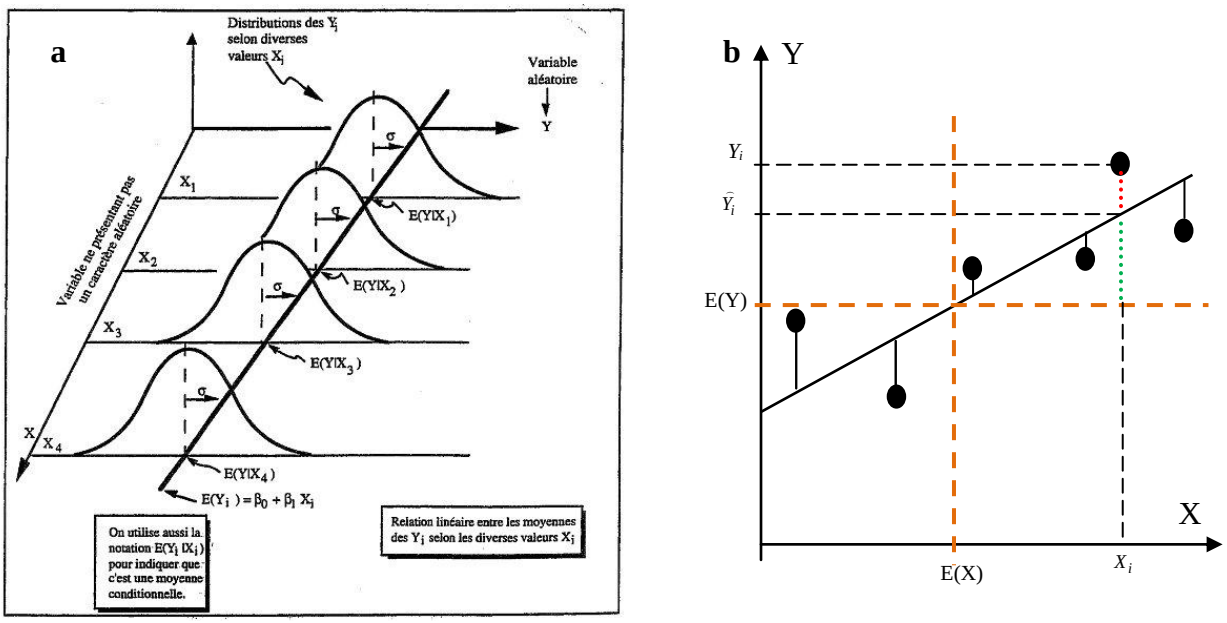

Figure 1. **a** Hypothèses de régression. **b**. Etablissement de la droite de régression $\hat{Y}$ sur le critère des moindres carrés. Notez que la droite passe par le point moyen (E(X), E(Y)) et la différence $(Y_i - E(Y)) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \mathrm{E}(Y))$ qui explique la décomposition de la variance (voir ci-dessous).

II RESOLUTION DU PROBLEME

Les paramètres β_0 (ordonnée à l'origine) et β_1 (pente) sont inconnus et doivent être estimés. Si les conditions 1 à 4 sont vérifiées, le *principe des moindres carrés ordinaires* (MCO) s'applique : on cherche les valeurs de β_0 et β_1 qui minimisent **la somme des carrés des résidus (SC_{res})** c'est-à dire que l'on minimise leur éparpillement (=dispersion) :

$$SC_{res} = \sum_i \left(y_i - (\beta_0 + \beta_1 \cdot x_i) \right)^2 .$$

Comme toujours, minimiser (respectivement maximiser) une expression $f(x)$ consiste à chercher la solution de $\frac{df(x)}{dx} = 0$, c'est-à-dire la valeur de x qui annule la dérivée de l'expression.

En dérivant cette expression par rapport à β_0 puis par rapport à β_1 puis en annulant ces dérivées on obtient deux équations dites **équations normales** :

$$n\,\beta_0 + \beta_1 \Sigma x_i = \Sigma y_i \qquad \text{et} \qquad \beta_0 \Sigma x_i + \beta_1 \Sigma x_i^2 = \Sigma\, y_i\, x_i$$

En résolvant ces équations on obtient les estimateurs des deux paramètres :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

où $\bar{x}$ et $\bar{y}$ sont les espérances (= moyennes) des n y_i et n x_i valeurs respectivement (parfois notées E(X) et E(Y)).

Notez bien que l'expression de β_1 n'est ni plus ni moins que : β_1= Cov(X,Y)/Var(X)

En examinant la figure **1b**, on peut se convaincre facilement que :

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$
$$SC_T = SC_E + SC_R$$

Avec :

SC_T = Variance totale (somme des carrés sur le total des données)
SC_E = Variance résiduelle (variance des résidus)
SC_R = Variance de la régression (calculée sur la distance de la régression à la moyenne E(Y))

On définit :
Le Carré Moyen de l'Erreur : $CM_E = SC_E/(n-2)$ et le Carré Moyen de la Régression : $CM_R = (SC_R / 1)$.
On peut montrer que :

$$E(CM_E) = \sigma^2 \text{ et } E(CM_R) = \sigma^2 + \beta_1^2 S_{xx}$$
$$S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

Donc, si l'hypothèse nulle est vraie, H0 : $\beta_1 = 0$ (Y et X **ne sont absolument pas corrélés**), alors les **Carrés Moyens CM_R et CM_E sont deux estimateurs sans biais de σ^2.**

Dans ces conditions, le rapport **$F = CM_R/CM_E$ est une variable aléatoire issue d'une loi de Fisher de degrés de liberté 1 et n-2** (la loi de Fisher admet 2 paramètres)**. On remarque alors que si H0 est vraie, alors $F \approx 1$**. Donc si F observé est grand (F >> 1), on a tout lieu de penser qu'il faut rejeter H0 !

III. RESULTATS

Les résultats issus de la décomposition de la variation de la variable réponse en somme de carrés sont fréquemment réunis sous la forme d'une table appelée table d'analyse de variance comme celle figurant dans la table ci-dessous. Il apparaît clairement ainsi que l'ANOVA n'est ni plus ni moins qu'une régression linéaire par la méthode des moindre carrés ordinaires et donc basée sur l'hypothèse d'homogénéité des variances des données et de normalité des mesures (2 contraintes d'application de cette méthode). Réduite à la comparaison de 2 moyennes (comme ici !) elle est équivalente à un test t de Student.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F
Régression	SC_R	1	$CM_R = SC_R/1$	CM_R/CM_E
Résidus	SC_E	n – 2	$CM_E = SC_R/(n - 2)$	
Total	SC_T	n – 1		

Table d'analyse de variance (ANOVA)

Dans le logiciel de statistique R, la table de l'analyse de variance est donnée en anglais :

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	p-value
Regression	1	SC_R	$CM_R = SC_R/1$	CM_R/CM_E	$P(F > f_{obs} \mid H_0)$
Residuals	n − 2	SC_E	$CM_E = SC_R/(n - 2)$		

Table d'analyse de variance (ANOVA)

IV DECISION FINALE

⇒La **p-valeur** (en anglais, **p-value**) **permet de prendre une décision quant au rejet ou non de la nullité du coefficient β_1** dans un modèle linéaire simple. Elle se lit : *Probabilité que F soit supérieure à Fobs sous l'hypothèse que H0 est vraie*. D'une façon plus intuitive, elle indique la quantité d'information manquante pour rejeter H1, que l'on accepte, alors que H0 est vraie malgré tout. En effet, si la p-valeur est inférieure au niveau de signification α spécifié par l'expérimentateur (par exemple α = 0.05), le test est dit statistiquement significatif à ce niveau α. L'hypothèse nulle H_0 est ainsi rejetée au niveau de signification α.

⇒ Le coefficient de détermination **R^2** exprime **en pourcentage** la part d'explication apportée par la variance de la régression à la variance totale :

$$R^2 = SC_R / SC_T$$

Ainsi, $R^2 = 0.8$ signifie que 80% de la variance totale des données sont expliquées par la régression. Le coefficient *r* de Pearson, variant entre -1 et 1 n'est ni plus ni moins que $r = \sqrt{R^2}$. Il est de moins en moins utilisé.

⇒Les tests d'auto-corrélation (hétéroscédaticité) sont aussi utilisés pour détecter la bonne dispersion autour de la régression des résidus.

⇒La normalité des résidus. La distribution des résidus doit être de type gaussienne. Des tests classiques de normalité (droite de Henry, test du X^2, aplatissement, symétrie…) peuvent être appliqués.

R : RAPPEL/PREMIER CONTACT

R est un environnement intégré d'un ensemble des softwares (logiciels) pour la manipulation de données, le traitement statistique de celles-ci, l'affichage graphique …et bien plus encore. R existe en version Windows et Linux (32/64) ainsi que Mac OS. Nous n'envisageons que la version Windows. Il y a 3 façons au moins de travailler avec R

1. En mode « console ». Ouvrir une console. Passer dans le répertoire de R (en principe C:\Programmes\R\bin) puis taper simplement R.
2. En mode Graphique. Clicker sur l'icône crée par l'installateur de R.

3. En utilisant l'interface Tinn-R. Une fois Tinn-R lancée, une icône ou bien à partir du menu R>Start/close connection>Rgui.

Remarquez bien que les modes 2 et 3 provoquent tous deux l'ouverture d'une fenêtre « R console ». La seule différence étant la présence de la fenêtre Tinn-R dans le dernier cas.
En mode 2 vous devrez entrer successivement vos commandes dans la console, ou mieux écrire ce que l'on appelle un script (Fichier > nouveau script), que l'on peut ensuite envoyer vers la console ligne à ligne (Ctrl + R) ou en intégralité.
En mode 3 vous pourrez préparer directement un « script » dans la fenêtre de Tinn-R, avec coloration syntaxique, puis au moyen d'un seul click envoyer vers R l'ensemble du script qui s'exécutera alors.
Tinn-R a bien d'autres fonctionnalités à découvrir…
Pour quitter R utiliser le menu de la console ou bien entrer la commande q() [les parenthèses indiquent que la commande q est en fait un appel de fonction].

On supposera désormais que vous êtes en mode 2 ou 3.
La première des choses est de spécifier le répertoire de travail courant (sauvegarde des scripts, fichiers de données et résultats). Utiliser le menu fichier de R et choisissez votre répertoire.

Obtenir de l'aide. Menu <Aide> : FAQ, Manuels et une aide en html très pratique avec une recherche automatique dans l'ensemble des packages de R du mot souhaité (essayez « **rnorm** » exemple). Autre façon : dans la console de R entrez la commande : « **help(rnorm**) » ou encore « **?rnorm** ». Essayer aussi « **?help** ». Si vous souhaitez avoir des informations sur un sujet, il suffit d'entrer quelque chose comme « **help.search("linear models")» .**

```
# Le texte suivant un '#' dans une ligne est interprété comme un commentaire par R et donc ignoré.
# R est sensible à la casse : 'A' est différent de 'a'.
# on essaie quelques commandes

# On génère deux vecteurs de nombre pseudo-aléatoires pour les coordonnées x et y.
x <- rnorm(50)
x                                        #voir ce que contient x
y <- rnorm(50)
y
x11(w=4, h=4)                            #Fixe la taille de la fenêtre
plot(x, y)
ls()
                                         # On essaie maintenant une autre représentation
colors()                                 # choix des couleurs avec la nomenclature exacte de R
hist(x,col="blue", border="green")
hist(y,col="Peachpuff3", border="tomato4")

rm(x,y)                                  # détruit des données  (vecteurs) x et y
ls()                                     # les objets x et y n'existent plus
graphics.off()                           # fermeture du graphique
```

Application. On mesure l'influence de la température sur la fonction de filtration de l'eau par une espèce de moule *Mytilus edulis*. 10 individus sont placés dans un litre d'eau colorée à une température donnée. On mesure toutes les 5 minutes la densité optique de l'eau (DO) qui caractérise la quantité de colorant non encore piégée par les animaux. Les données sont situées sur un serveur à l'université de Lyon I, sous la forme d'un fichier texte.

```
resu <- read.table("http://pbil.univ-lyon1.fr/R/donnees/moules.txt",sep = "\t", header = TRUE)

# affichez le contenu de resu. Que contient-il ?
# on sauve sur le disque local pour une utilisation ultérieure
write.table(resu, append = FALSE, file = "mytilus.txt", sep="\t", col.names=TRUE)

#on entre maintenant les donnees  « températures »
temp <- c(7.5, 15, 22, 27, 34, 15, 18, 22, 25, 28)

x  <-  resu[1,2:10]
x                  # contenu de x
x <- resu[ ,2]     # A comparer avec resu[2, ]
```

```
x <- resu[ ,1]      # on entre maintenant le temps dans la variable x
y <- resu[ ,2]      # puis en y les valeurs dépendantes

A <- lm(y~x)                 # une régression lineaire
A                            # le contenu de A
summary(A)                   # résultats statistiques ; que pensez-vous ?

# NB : On aurait pu directement écrire : A <- summary(lm(resu[ ,2]~resu[ ,1]))

plot(x,y)                          # affichage du nuage de dispersion
abline(coef(A), col="red")         # la droite de regression

new <- data.frame( x=seq(0,58,length=20) ) # on tente des predictions
p <- predict(A, new)
points( p ~ new$x, type='l' )      # on definit le type d'affichage

          #intervalle de confiance de la droite Y = aX + b
p <- predict(A, new, interval='confidence' )
points( p[,2] ~ new$x, type='l', col="green" )
points( p[,3] ~ new$x, type='l', col="green" )

          #intervalle de prédiction de E[Y|X=x]
p <- predict(A, new, interval='prediction')
points( p[,2] ~ new$x, type='l', col="red"  )
points( p[,3] ~ new$x, type='l', col="red"  )

B <- residuals(A)             # les résidus de la régression
summary(B)                    # notez la moyenne
var(B) ; sqrt(var(B))         # variance et écart-type
plot(B)                       # affichage

summary(aov(A))               # une analyse de variance de la réponse de y sur  x
                              # Comparez avec le résultat fourni par la régression linéaire

sink(file="A.txt")            # pour terminer on sauve le résultat de la régression
sink()
q()                           # et on quitte R
```

APPENDICE :
LES RESULTATS DE LA REGRESSION LINEAIRE ET DE L'ANALYSE DE VARIANCE DE L'EXEMPLE « MOULES.TXT »

1. Régression linéaire

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-19.182  -12.682  -3.055   8.818  34.655

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  468.9273    11.2840    41.56   1.35e-11 ***
x             -5.6582     0.3327   -17.00   3.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.45 on 9 degrees of freedom
Multiple R-squared: 0.9698,     Adjusted R-squared: 0.9665
```

2. ANOVA

```
          Df  Sum Sq  Mean Sq  F value    Pr(>F)
x          1  88041    88041   289.15  3.779e-08 ***
Residuals  9   2740      304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```