Fiche TD avec le logiciel R : tdr33

# Analyses de la variance

D. Chessel & A.B. Dufour

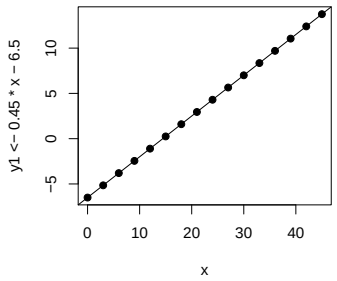
Bases du modèle linéaire à effet fixe : régression simple, analyse de variance, régression multiple, analyse de covariance

## Table des matières

# 1 Introduction

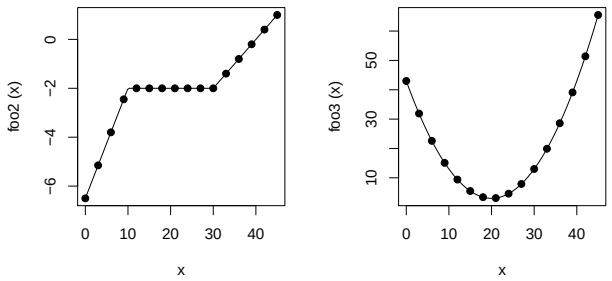
L'essentiel, pour approcher le modèle linéaire, est de comprendre sa définition. On supposera ensuite que la réalité est faite, à peu près, comme ça. Ce n'est pas vrai mais ce peut être utile.

```
x <- seq(from = 0, to = 45, by = 3)
plot(x, y1 <- 0.45 * x - 6.5, pch = 20, cex = 1.5)
abline(-6.5, 0.45)
```



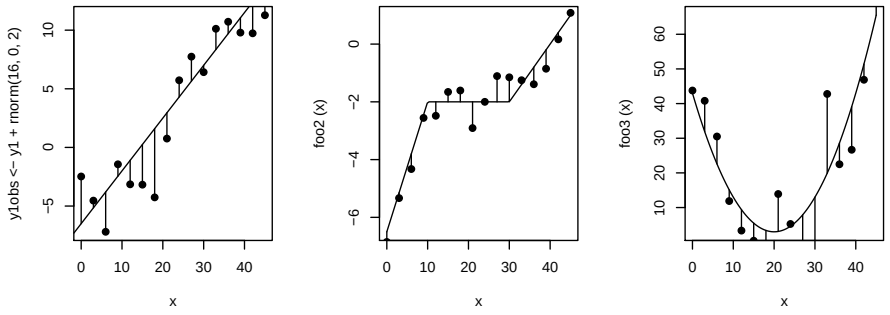

Ceci semble un modèle linéaire bien propre, parce que les points sont sur une droite. C'est un modèle linéaire bien propre, certes, mais pas à cause de la droite. Voilà deux modèles linéaires sans droite :

```
par(mfrow = c(1, 2))
foo2 <- function(x) {
    (0.45 * x - 6.5) * (x < 10) + ((x >= 10) & (x < 30)) * (-2) +
        (x >= 30) * (0.2 * x - 8)
}
plot(foo2, 0, 45)
points(x, y2 <- foo2(x), pch = 20, cex = 1.5)
foo3 <- function(x) {
    (x - 20)^2/10 + 3
}
plot(foo3, 0, 45)
points(x, y3 <- foo3(x), pch = 20, cex = 1.5)
```



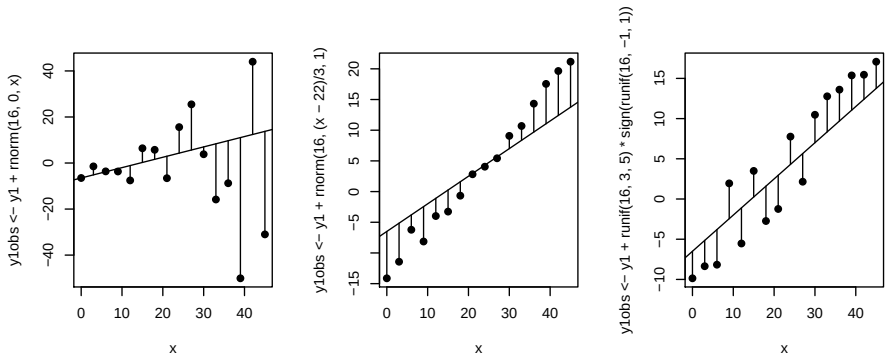

Ce qui fait un modèle linéaire est qu'on additionne des composantes d'une part, on y reviendra, et que les erreurs autour du modèle sont normales, de variance constante et indépendantes. Voilà 3 vrais modèles linéaires :

```
par(mfrow = c(1, 3))
plot(x, y1obs <- y1 + rnorm(16, 0, 2), pch = 20, cex = 1.5)
abline(c(-6.5, 0.45))
segments(x, y1, x, y1obs)
plot(foo2, 0, 45)
points(x, y2obs <- y2 + rnorm(16, 0, 0.5), pch = 20, cex = 1.5)
segments(x, y2, x, y2obs)
plot(foo3, 0, 45)
points(x, y3obs <- y3 + rnorm(16, 0, 10), pch = 20, cex = 1.5)
segments(x, y3, x, y3obs)
```




Voilà des modèles linéaires qui n'en sont pas. Les erreurs du premier ne sont pas de variance constante, les erreurs du second ne sont pas de moyenne nulle, les erreurs du troisième ne sont pas distribuées selon une loi normale.

```
par(mfrow = c(1, 3))
plot(x, y1obs <- y1 + rnorm(16, 0, x), pch = 20, cex = 1.5)
abline(c(-6.5, 0.45))
segments(x, y1, x, y1obs)
plot(x, y1obs <- y1 + rnorm(16, (x - 22)/3, 1), pch = 20, cex = 1.5)
abline(c(-6.5, 0.45))
segments(x, y1, x, y1obs)
plot(x, y1obs <- y1 + runif(16, 3, 5) * sign(runif(16, -1, 1)),
    pch = 20, cex = 1.5)
abline(c(-6.5, 0.45))
segments(x, y1, x, y1obs)
```




# 2 Les objets de la classe lm

La fonction fondamentale s'appelle simplement `lm` (*linear model*).

Soit un modèle linéaire très simple formé de l'**explicative** `x`, d'un *vrai* **modèle** `ymod` (ça n'existe que dans les fiches de TD, mais il faut bien s'instruire) et d'une variable **observée** `yobs`. L'observation est le modèle perturbé par une erreur.

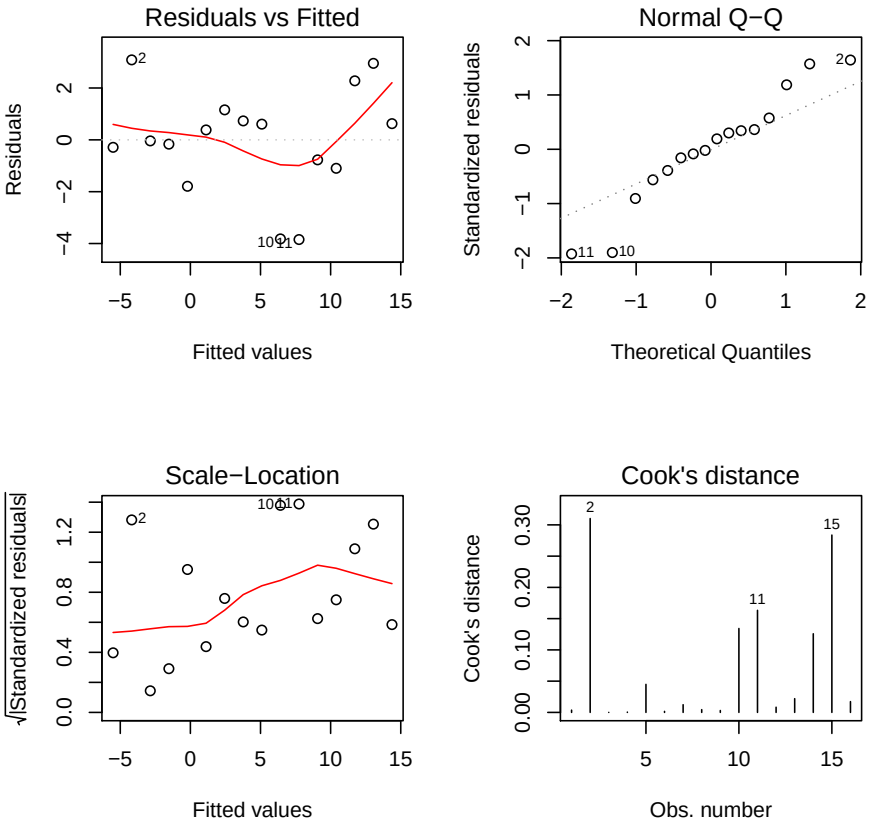
```
x <- seq(from = 0, to = 45, by = 3)
ymod <- 0.45 * x - 6.5
yobs <- ymod + rnorm(16, 0, 2)
yobs <- signif(yobs, 2)
x
 [1]  0  3  6  9 12 15 18 21 24 27 30 33 36 39 42 45
ymod
 [1] -6.50 -5.15 -3.80 -2.45 -1.10  0.25  1.60  2.95  4.30  5.65  7.00  8.35  9.70
[14] 11.05 12.40 13.75
yobs
 [1] -5.8 -1.1 -2.9 -1.7 -2.0  1.5  3.6  4.5  5.7  2.6  3.9  8.3  9.3 14.0 16.0 15.0
lm1 <- lm(yobs ~ x)
```

Questions :

1. Quelle est la classe de l'objet `lm1` ?
2. `lm1` est-il une liste ? *Les classes sont emboîtées*
3. Combien l'objet `lm1` a-t-il de composantes ?
4. Quels sont les noms des composantes ?
5. Que contient la composante `call` ? A quoi ça sert ? *Essayez étonnant, non ?*
6. Que contient la composante `coefficients` ? A quoi ça sert ?
7. Que contient la composante `residuals` ? A quoi ça sert ?
8. Quelle est la relation entre les deux objets suivants ?

   ```
   model.matrix(~x)%*%lm1$coefficients
   lm1$fitted.values
   ```

```
par(mfrow = c(2, 2))
plot(lm1, 1:4)
```




Questions :

1. Que se passe-t-il quand on ne prévoit pas de multifenêtrage ?
2. Où retrouver les coordonnées du point étiqueté 2 dans la première fenêtre ? (Réponse : -4.186, 3.086)
3. Où retrouver les coordonnées du point étiqueté 2 dans la troisième fenêtre ? (Réponse : -4.186, 1.3277)

La seconde fenêtre contient un graphique quantile-quantile normal des résidus (normalité des résidus). Noter que chacun des graphiques proposés est issu d'une recherche approfondie. Le qq-plot est de Wilk & Gnanadesikan [10]. Il est validé par Cleveland [3, p. 143]. Les modes de lecture sont décrits dans des ouvrages célèbres comme ceux de Tuckey [9], du Toit & al. [6, voir p. 49], Chambers & al. [1] et Cleveland [2].

Le dernier graphe est celui des distances de Cook. Il donne pour chacun des points de mesure la distance entre les paramètres estimés par la régression avec et sans ce point. Si l'importance du rôle de chaque point est concentré sur quelques valeurs, la régression n'est pas bonne (prise en compte de points aberrants). Voir Cook & Weisberg [4].

# 3 Variances d'échantillonnage

On peut refaire l'expérience.

1. Définir la variable explicative ;
2. Générer une erreur normale ;
3. Additionner les deux.

```
x <- seq(0, 45, by = 5)
e <- rnorm(10, 0, 8)
y <- x + e
plot(x, y)
abline(0, 1, lwd = 2)
lmparori <- lm(y ~ -1 + x)
abline(lmparori, lwd = 2, lty = 2, col = "red")
```

Question : quelles sont les possibilités de paramétrer une droite à rajouter sur un graphe ?

L'exercice est fondamental. Il construit les données conformément à un modèle. Une valeur de y est la réalisation d'une variable aléatoire gaussienne de moyenne $\mu$ et de variance $\sigma^2$. $\mu$ est une fonction de $x$ : ici la plus simple qui soit $\mu = x$. On écrit en général $E(Y) = ax$ (la moyenne est modélisée) et $Var(Y) = Cte = \sigma^2$ (la variance est constante). L'erreur est gaussienne. Faire la régression, c'est estimer les valeurs inconnues $(a, \sigma)$ à partir de l'échantillon dans ce type de modèle (trop beau pour être " biologique " ?)

Entre le modèle vrai (pour avoir un modèle vrai, il vaut mieux le construire soi même) et l'estimation, il y a évidemment la variance d'échantillonnage, c'est à dire le rôle du hasard sur les observations qui par essence sont aléatoires autour d'une valeur prévue. Et évidemment, tout ce qu'on observe est faux (enfin, assez faux, mais pas trop). On a mis $a = 1$ et on a estimé 1.2401. On a mis $\sigma = 8$ et on a trouvé 4.5441.

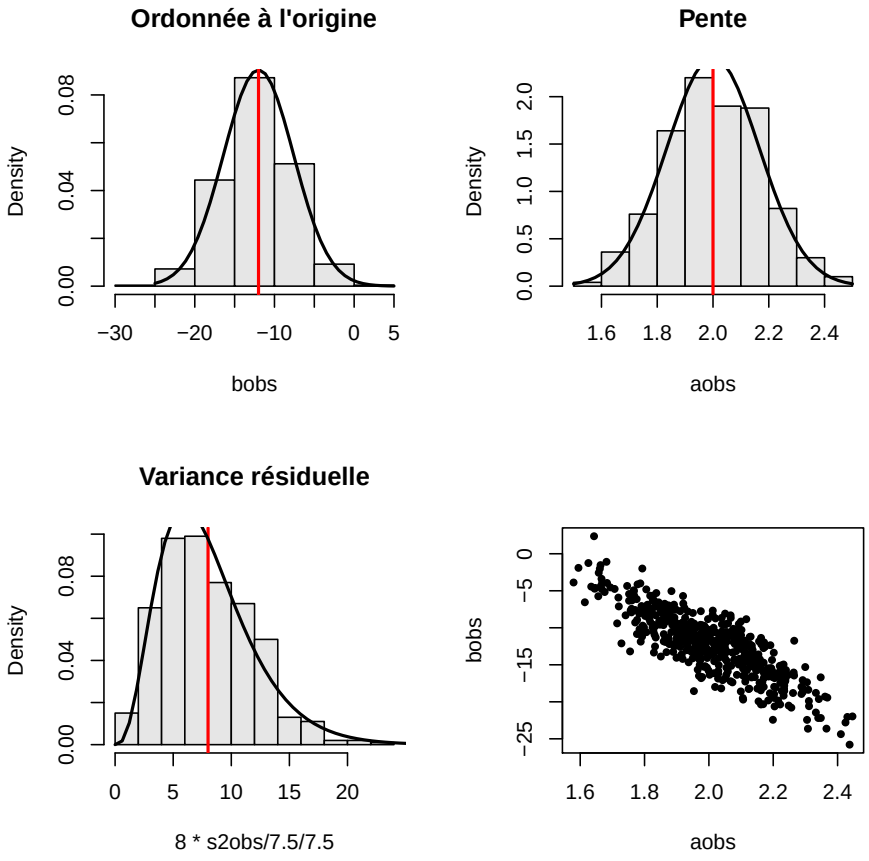
```
summary(lmparori)
```

```
Call:
lm(formula = y ~ -1 + x)
Residuals:
    Min      1Q  Median      3Q     Max
-9.5733 -2.6401 -0.5186  1.0084  6.4110

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  1.24005    0.05383   23.04 2.60e-09 ***
---
Signif. codes:  0

Residual standard error: 4.544 on 9 degrees of freedom
Multiple R-squared: 0.9833,        Adjusted R-squared: 0.9815
F-statistic: 530.6 on 1 and 9 DF,  p-value: 2.603e-09
```

Il est alors, une fois au moins, utile d'examiner la variance d'échantillonnage. Supposons que le vrai modèle soit $E(Y) = ax + b$ avec $a = 2$, $b = -12$ et $\sigma = 7.5$. On a toujours $n = 10$. Faisons 500 fois l'expérience :

```
fun1 <- function(k) {
    y <- 2 * x - 12 + rnorm(10, 0, 7.5)
    lm1 <- lm(y ~ x)
    return(c(lm1$coefficients, summary(lm1)$sigma))
}
echa <- matrix(sapply(1:500, fun1), ncol = 3, byr = T)
par(mfrow = c(2, 2))
det <- (10 * sum(x^2) - sum(x)^2)
bobs <- echa[, 1]
```

```
sdb <- 7.5 * sqrt(sum(x^2)/det)
hist(bobs, main = "Ordonnée à l'origine", proba = T, col = grey(0.9))
bx <- seq(-25, 5, le = 40)
lines(bx, dnorm(bx, -12, sdb), lwd = 2)
abline(v = -12, lwd = 2, col = "red")
aobs <- echa[, 2]
ax <- seq(1.5, 2.5, le = 40)
sda <- 7.5 * sqrt(10/det)
hist(aobs, main = "Pente", proba = T, col = grey(0.9))
lines(ax, dnorm(ax, 2, sda), lwd = 2)
abline(v = 2, lwd = 2, col = "red")
s2obs <- echa[, 3]^2
hist(8 * s2obs/7.5/7.5, proba = T, main = "Variance résiduelle",
    col = grey(0.9))
abline(v = 8, lwd = 2, col = "red")
sx <- seq(0, 30, le = 50)
lines(sx, dchisq(sx, df = 8), lwd = 2)
plot(aobs, bobs, pch = 20)
```




Cette figure est une illustration expérimentale d'un ensemble de théorèmes fondamentaux. Les valeurs du prédicteur $x_1, x_2, \cdots, x_n$ sont fixées. La réponse est une variable aléatoire $Y_i$ qui suit une loi normale de moyenne $ax_i + b$ et de variance $\sigma^2$. Pour une expérience, c'est-à-dire un tirage aléatoire simple de la variable $Y_1, \cdots, Y_n$, **l'estimateur** de l'ordonnée à l'origine est une variable aléatoire $A$ qui à chaque expérience donne une valeur numérique, ce qu'on appelle une **estimation**. Nous avons reproduit par un échantillon de 1000 valeurs

estimées la loi de l'estimateur. Sa moyenne est $a$ la vraie valeur. On dit que l'estimateur est juste ou **sans biais**.

En moyenne, je trouve la vraie valeur. C'est bien là le drame. C'est en moyenne et je ne ferai qu'un essai. Autant dire que la variance joue un rôle crucial, parce que si elle est grande, que la moyenne soit correcte implique que mon unique valeur peut être totalement fausse. Si au contraire elle est petite, je serai vraisemblablement pas loin de la vraie valeur. Un estimateur de petite variance est **précis**.

Notre simulation illustre bien que sous les hypothèses énoncées l'estimateur $A$ suit une loi normale de moyenne $a$ et de variance :

$$Var(A) = \sigma^2 \frac{n}{n\sum_i(x_i^2) - (\sum_i(x_i))^2}$$

De même, à chaque expérience, l'estimateur de l'ordonnée à l'origine qui est une variable aléatoire $B$ donne une estimation $\hat{b}$. Sous les hypothèses énoncées, l'estimateur $B$ suit une loi normale de moyenne $b$, il est sans biais, et de variance :

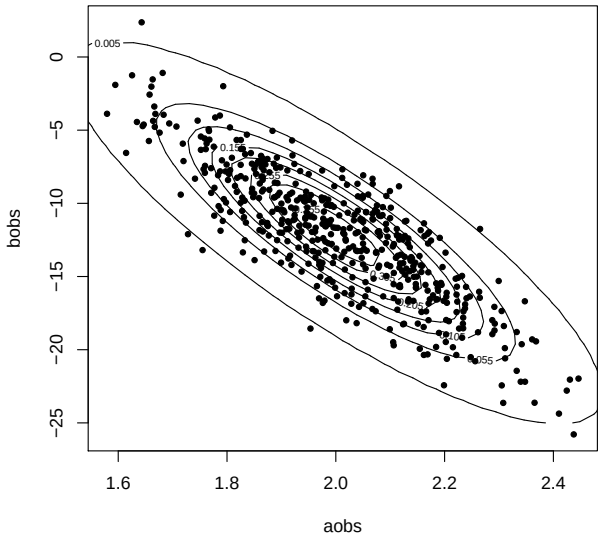
$$Var(B) = \sigma^2 \frac{\sum_i(x_i^2)}{n\sum_i(x_i^2) - (\sum_i(x_i))^2}$$

Il est de moindre intérêt de connaître les formules pour calculer les estimations ou les variances d'estimation dès que le logiciel prend en charge ces calculs et les exécutent en général correctement. On peut donc remplacer le temps nécessaire au calcul par un peu de réflexion. En particulier, sur la question comment diminuer la variance d'échantillonnage ? La réponse est simple : en augmentant $n$ (ça c'est bien connu, il n'y a jamais assez de données) mais aussi en augmentant la variance du prédicteur. Et comment augmenter cette variance, tout en restant dans un domaine raisonnable ? Réponse : en mettant une moitié des valeurs au minimum et une moitié des valeurs au maximum. Et pourquoi on ne fait pas ça, si on sait que c'est ce qu'il y a de mieux ? Réponse parce qu'on n'est jamais certain que le modèle linéaire soit le bon. C'est le meilleur dans ce cas, mais dans ce cas seulement.

Pour le troisième paramètre, il en est de même. A chaque expérience, l'estimateur $V$ de $\sigma^2$ donne une estimation $\hat{\sigma}^2$. Nous illustrons le fait que la loi de la variable aléatoire $\frac{(n-2)A}{\sigma^2}$ suit une loi $\chi^2$ à $n-2$ degrés de liberté, ce qui introduit à l'analyse de la variance.

Enfin, les estimateurs sur les deux premiers paramètres ne sont pas indépendants, en ce sens que l'erreur qu'on commet sur le premier est liée à l'erreur qu'on commet sur le second. On peut préciser :

```
plot(aobs, bobs, pch = 20)
covab <- 7.5 * 7.5 * (-sum(x)/det)
xyg <- expand.grid(x = ax, y = bx)
library(mvtnorm)
z <- dmvnorm(xyg, me = c(2, -12), sigma = matrix(c(sda^2, covab,
    covab, sdb^2), 2))
z <- matrix(z, 40, 40)
contour(ax, bx, z, add = T, lev = seq(0.005, 0.5, by = 0.05))
```




Un malheur n'arrive jamais seul. Si l'erreur commise sur un des paramètres est grande et positive, elle sera grande et négative sur l'autre. Plus précisément, les deux estimateurs suivent conjointement une loi de Gauss bivariée de covariance égale à :

$$Cov(A, B) = \sigma^2 \frac{-\sum_i (x_i)}{n \sum_i (x_i^2) - (\sum_i (x_i))^2}$$

Dans toute la mesure du possible, on la gardera présente à l'esprit, quand on fait un modèle linéaire, l'idée que tous les résultats sont faux et, par là même, sans grand intérêt s'ils ne sont pas accompagnés de précisions sur . . .la précision.

# 4 La régression est-elle dangereuse ?

On peut reprendre le chapitre 9 dans [8, p.175] :

```
lm(c(3, 10) ~ -1 + c(2, 3))
Call:
lm(formula = c(3, 10) ~ -1 + c(2, 3))
Coefficients:
c(2, 3)
  2.769
anova(lm(c(3, 10) ~ -1 + c(2, 3)))
Analysis of Variance Table
Response: c(3, 10)
          Df Sum Sq Mean Sq F value Pr(>F)
c(2, 3)    1 99.692  99.692  10.711 0.1888
Residuals  1  9.308   9.308
```

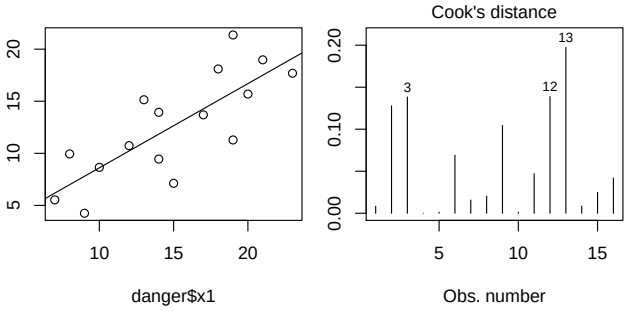
Pour retrouver les détails des calculs décrits dans [8, p.183] :

```
t <- c(3, 3, 6, 10, 10, 12, 15, 18, 20)
x <- c(7, 7, 6, 8, 8, 7, 5, 4, 3)
y <- c(39.2, 37.8, 35.8, 51.2, 47.4, 45.2, 39.7, 37.4, 35.1)
lmt1 <- lm(y ~ -1 + t + x)
lmt1
```

```
predict(lmt1)
residuals(lmt1)
summary(lmt1)
```

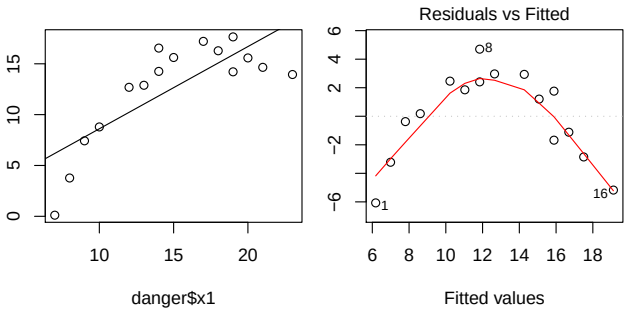
Partir du tableau 1.1 dans [7, p. 22] :

```
danger <- read.table("http://pbil.univ-lyon1.fr/R/donnees/danger.txt",
    header = TRUE)
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
plot(danger$x1, danger$y1)
lm1 <- lm(danger$y1 ~ danger$x1)
abline(lm1)
plot(lm1, 4)
```



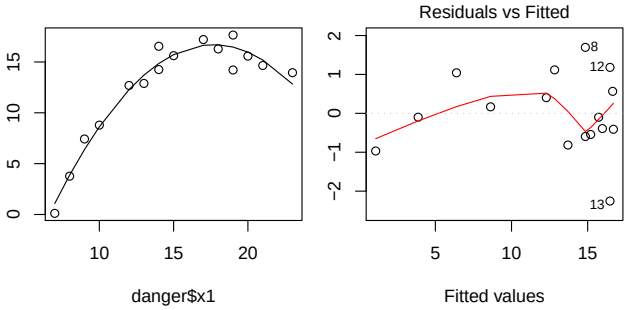

*Une bonne régression*

```
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
plot(danger$x1, danger$y2)
lm2 <- lm(danger$y2 ~ danger$x1)
abline(lm2)
plot(lm2, 1)
```



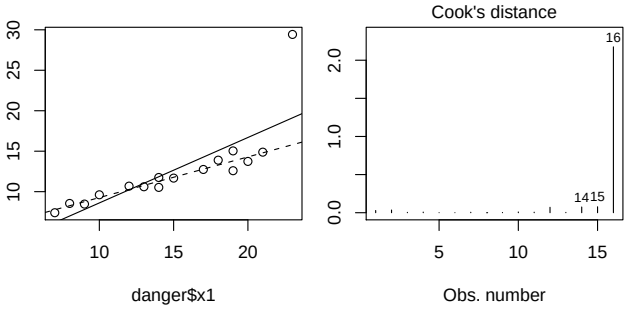

*Résidus autocorrélés*

```
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
lm2po <- lm(y2 ~ poly(x1, 2), data = danger)
plot(danger$x1, danger$y2)
lines(danger$x1, predict(lm2po))
plot(lm2po, 1)
```



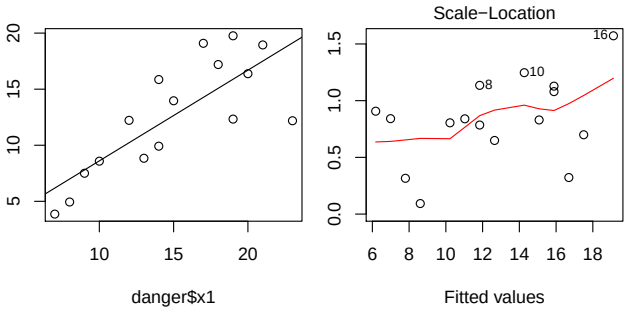

*Régression polynomiale*

```
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
plot(danger$x1, danger$y3)
lm3 <- lm(danger$y3 ~ danger$x1)
abline(lm3)
abline(lm(danger$y3[1:15] ~ danger$x1[1:15]), lty = 2)
plot(lm3, 4)
```




*Point pivot*

```
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
plot(danger$x1, danger$y4)
lm4 <- lm(danger$y4 ~ danger$x1)
abline(lm4)
plot(lm4, 3)
lm4 <- lm(y4 ~ x1, data = danger)
```




*Variance non constante*

```
coefficients(lm1)
coefficients(lm2)
coefficients(lm3)
coefficients(lm4)
```

# 5 Analyse de variance

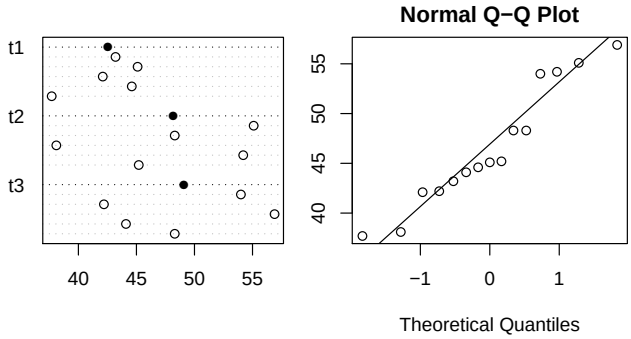
## 5.1 Un facteur contrôlé

Exercice proposé par P. Dagnelie [5, Exercice 14.1 p. 102] *Quinze veaux ont été répartis au hasard en trois lots, alimentés chacun d'une façon différente. Les gains de poids observés au cours d'une même période et exprimés en kg étant les suivants, peut-on admettre qu'il n'y a pas de relation entre l'alimentation et la croissance des veaux ?*

| | t1 | t2 | t3 |
|---|---|---|---|
| 1 | 37.70 | 45.20 | 48.30 |
| 2 | 44.60 | 54.20 | 44.10 |
| 3 | 42.10 | 38.10 | 56.90 |
| 4 | 45.10 | 48.30 | 42.20 |
| 5 | 43.20 | 55.10 | 54.00 |

Présenter les données sous la forme du lien entre un facteur et une réponse :

```
ali <- rep(c("t1", "t2", "t3"), c(5, 5, 5))
par(mfrow = c(1, 2))
par(mar = c(4, 2, 2, 1))
is.factor(ali)
is.character(ali)
ali <- as.factor(ali)
is.factor(ali)
levels(ali)
gain <- c(37.7, 44.6, 42.1, 45.1, 43.2, 45.2, 54.2, 38.1, 48.3,
    55.1, 48.3, 44.1, 56.9, 42.2, 54)
mgain <- tapply(gain, ali, mean)
dotchart(gain, gr = ali, gdata = mgain, gpch = 16)
anova(lm(gain ~ ali))
qqnorm(gain)
qqline(gain)
```

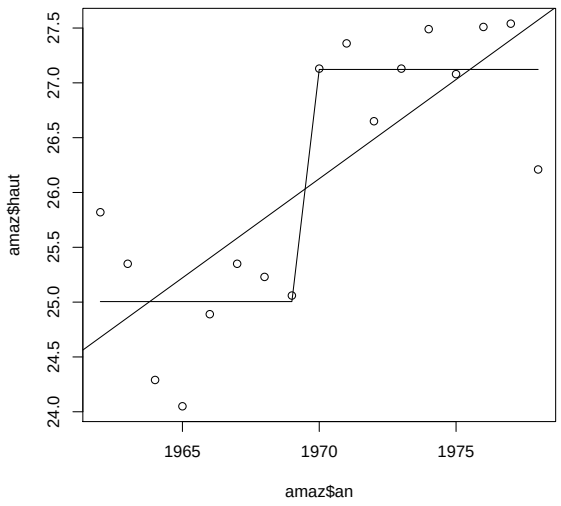



Commenter chaque ligne de commande, identifier l'intention et le résultat. *Le test est non significatif et l'ensemble des données peut être considéré comme*

*un échantillon aléatoire simple d'une loi normale* vous semble-t-il un résumé acceptable ? A commenter. Où est le danger ? Une situation particulièrement favorable ? En quoi ?

## 5.2 Unité entre analyse de variance et régression simple

| | an | haut | bas |
|---|---|---|---|
| 1 | 1962.00 | 25.82 | 18.24 |
| 2 | 1963.00 | 25.35 | 16.50 |
| 3 | 1964.00 | 24.29 | 20.26 |
| 4 | 1965.00 | 24.05 | 20.97 |
| 5 | 1966.00 | 24.89 | 19.43 |
| 6 | 1967.00 | 25.35 | 19.31 |
| 7 | 1968.00 | 25.23 | 20.85 |
| 8 | 1969.00 | 25.06 | 19.54 |
| 9 | 1970.00 | 27.13 | 20.49 |
| 10 | 1971.00 | 27.36 | 21.91 |
| 11 | 1972.00 | 26.65 | 22.51 |
| 12 | 1973.00 | 27.13 | 18.81 |
| 13 | 1974.00 | 27.49 | 19.42 |
| 14 | 1975.00 | 27.08 | 19.10 |
| 15 | 1976.00 | 27.51 | 18.80 |
| 16 | 1977.00 | 27.54 | 18.80 |
| 17 | 1978.00 | 26.21 | 17.57 |




Un des aspects les plus frappants de l'hydrologie de la haute Amazone est la fluctuation saisonnière marquée du niveau des eaux. Les niveaux annuels des hautes et basses eaux ont été relevés de 1962 à 1978 à Iquitos. haut =

Hautes eaux (m). bas = Basses Eaux (m). A partir de 1970, l'ouverture de routes dans la haute vallée de l'Amazone a autorisé une déforestation à large échelle. Cette pratique est susceptible d'avoir des conséquences climatologiques et hydrologiques importantes. Ces conséquences sont-elles perceptibles dans les données ci-dessus ? Télécharger le fichier :

`http://pbil.univ-lyon1.fr/R/donnees/amaz.txt`.

```
anova(lm(haut ~ an, data = amaz))
plot(amaz$an, amaz$haut)
abline(lm(haut ~ an, data = amaz))
lines(amaz$an, predict(lm(haut ~ an >= 1970, data = amaz)))
anova(lm(haut ~ an >= 1970, data = amaz))
```

*Apporter des arguments pour choisir un modèle. Noter que l'assertion* " A partir de 1970, l'ouverture de routes ... " *justifie* ***à priori*** *l'usage de l'hypothèse alternative* " avant et après sont différents ".

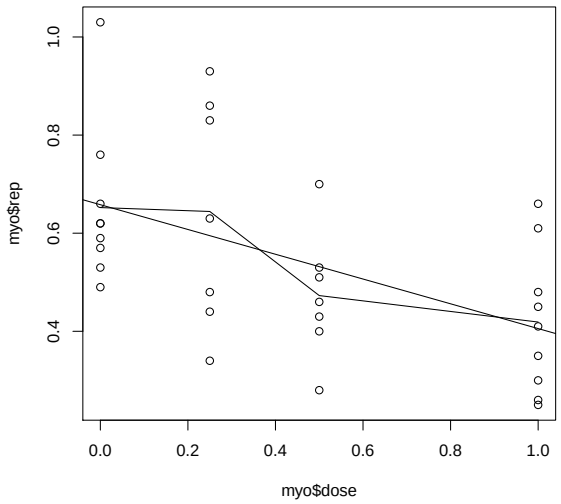
## 5.3 Ouabaïne et noradrénaline

Chez le rat, on teste l'effet de l'ouabaïne sur la teneur en noradrénaline du myocarde. Les résultats sont dans le tableau ci-dessous. On note x la dose d'ouabaïne injectée et y la teneur en noradrénaline.

| Ouabaïne (mg/kg) | | | |
|---|---|---|---|
| **0** | **0.25** | **0.5** | **1** |
| 0.49 | 0.63 | 0.51 | 0.66 |
| 0.66 | 0.93 | 0.53 | 0.48 |
| 0.59 | 0.48 | 0.28 | 0.25 |
| 0.62 | 0.34 | 0.7 | 0.3 |
| 0.76 | 0.83 | 0.43 | 0.35 |
| 0.57 | 0.44 | 0.4 | 0.61 |
| 0.62 | 0.86 | 0.46 | 0.45 |
| 0.53 | | | 0.26 |
| 1.03 | | | 0.41 |

Retrouver ces données dans :

`"http://pbil.univ-lyon1.fr/R/donnees/myo.txt`.



```
plot(myo$dose, myo$rep)
abline(lm(rep ~ dose, data = myo))
dose.fac <- as.factor(myo$dose)
lines(myo$dose, predict(lm(myo$rep ~ dose.fac)))
lm1 <- lm(myo$rep ~ myo$dose)
lm2 <- lm(myo$rep ~ dose.fac)
anova(lm1)
anova(lm2)
summary(lm1)
summary(lm2)
anova(lm2, lm1)
anova(lm1, lm2)
```

*Rappeler pourquoi la comparaison de ces deux modèles est valide contrairement au cas précédent sur l'Amazone. Donner un nom au dernier test effectué et résumer l'information acquise.*

## 5.4 Un exemple pour rire

On a trouvé dans un vieux polycopié d'exercices de statistique pour biologistes cet énoncé :

> On a mesuré dans 4 régions fixées, les poids en kg de cerfs mâles. Les régions A et C sont situées dans le nord de la France, la région B à l'est et la région D à l'ouest.

| | rA | rB | rC | rD |
|---|---|---|---|---|
| 1 | 60.50 | 72.10 | 62.00 | 40.10 |
| 2 | 62.10 | 70.70 | 60.30 | 36.50 |
| 3 | 57.30 | 72.50 | 57.50 | 39.70 |
| 4 | 55.00 | 68.00 | 61.80 | 42.30 |
| 5 | 64.20 | 67.40 | 62.50 | 45.70 |
| 6 | 61.10 | 72.60 | 61.20 | 41.40 |
| 7 | 60.00 | 67.20 | 64.50 | 40.60 |
| 8 | 59.70 | 68.90 | 56.30 | 39.80 |
| 9 | 60.20 | 74.20 | 63.10 | 42.00 |
| 10 | 59.90 | 71.40 | 60.80 | 42.90 |

Retrouver ces données à :

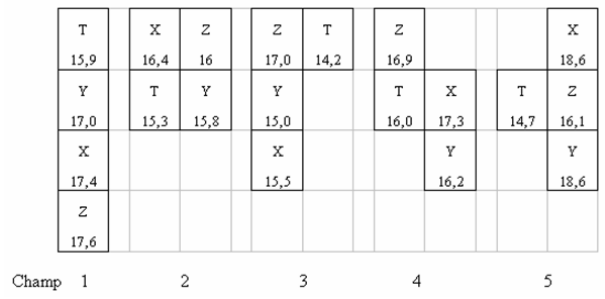
`http://pbil.univ-lyon1.fr/R/donnees/cerf.txt`.

```
bartlett.test(cerf$poi, cerf$reg)
```

Expliquer en quoi ce résultat montre qu'il s'agit de données fabriquées et expliquer la raison qui pouvait pousser vos ancêtres à commettre de tels méfaits.

## 5.5 Deux facteurs

On veut tester l'efficacité de trois insecticides X, Y et Z contre la pyrale (papillon) du maïs. 5 champs non contigus de forme variable sont subdivisés en 4 parcelles de même surface et sur chacune on teste un insecticide ou rien (témoin T). Pour chaque parcelle, on mesure le poids de grains en kg d'une nombre constant de plants.




Appeler `champ` et `prod` les facteurs contrôlés et `rep` la réponse. Implanter les données.

1. Comparer les modèles :

```
lm(rep ~ prod+champ)  ET lm(rep ~ champ+prod) ;
```

2. Contrôler les résidus ;
3. Identifier et comparer leurs coefficients ;
4. Comparer leurs prédictions ;
5. Retrouver manuellement leurs prédictions à partir de leur *summary* ;
6. Comparer leur *summary* ;
7. Comparer leurs *anova* ;

Enlever une seule valeur (supposer par exemple qu'un prédateur ait ruiné les plants subissant le traitement T dans le champ 3, rien de plus ordinaire) et recommencer l'exercice. Discuter l'intérêt des plans orthogonaux. Pourquoi portent-ils ce nom ?

## 5.6 Interaction

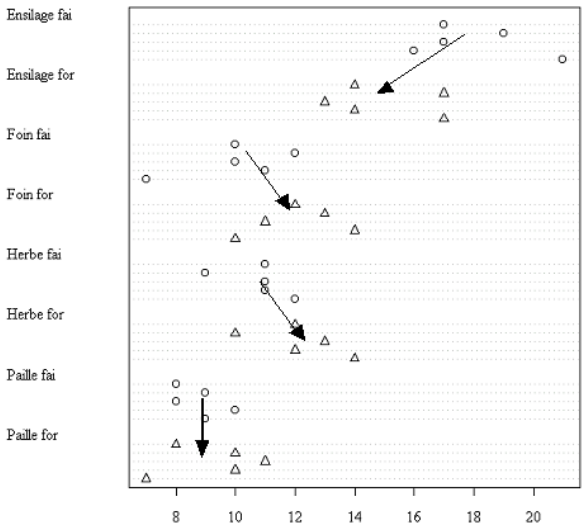
Dans une expérience sur le rendement des vaches laitières, on a choisi 40 animaux aussi identiques que possible et on les a répartis en 8 groupes de 5. Chaque groupe a été soumis à une alimentation différente.

| | Aliment | | | |
|---|---|---|---|---|
| Dose | Paille | Foin | Herbe | Ensilage |
| | 8 | 12 | 12 | 14 |
| | 10 | 13 | 10 | 17 |
| forte | 11 | 11 | 13 | 13 |
| | 10 | 14 | 12 | 14 |
| | 7 | 10 | 14 | 17 |
| | 8 | 10 | 11 | 17 |
| | 9 | 12 | 9 | 19 |
| faible | 8 | 10 | 11 | 17 |
| | 10 | 11 | 11 | 16 |
| | 9 | 7 | 12 | 21 |

Retrouver ces données à :

`http://pbil.univ-lyon1.fr/R/donnees/vache.txt`.

Retrouver le graphe ci-dessous. Quel est le nom du phénomène mis en évidence. Quelle est sa signification statistique. Etudier et donner un sens aux coefficients du modèle `lm(rep ali*dose,data=vache)`.




# 6 Régression multiple

Les données suivantes concernant 20 villes sont extraites d'une étude sur la pollution atmosphérique des villes des États-Unis : On a noté :

| | pol | tem | usi | pop |
|---|---|---|---|---|
| Atlanta | 24 | 62 | 368 | 497 |
| Baltimore | 47 | 55 | 625 | 905 |
| Chicago | 110 | 51 | 3344 | 3369 |
| Denver | 17 | 52 | 454 | 515 |
| Des Moines | 17 | 49 | 104 | 201 |
| Detroit | 35 | 50 | 1064 | 1513 |
| Hartford | 56 | 49 | 412 | 158 |
| Indianapolis | 28 | 52 | 361 | 746 |
| Jacksonville | 14 | 68 | 136 | 529 |
| Kansas City | 14 | 55 | 381 | 507 |
| Little Rock | 13 | 61 | 91 | 132 |
| Louisville | 30 | 56 | 291 | 593 |
| Miami | 10 | 76 | 207 | 335 |
| Minneapolis | 29 | 44 | 669 | 744 |
| New Orleans | 9 | 68 | 204 | 361 |
| Phoenix | 10 | 70 | 213 | 582 |
| San Francisco | 12 | 57 | 453 | 716 |
| Washington | 29 | 57 | 434 | 757 |
| Wichita | 8 | 57 | 125 | 277 |
| Wilmington | 36 | 54 | 80 | 80 |

- `pol` : teneur annuelle moyenne de l'air en SO2 en mg/m3
- `tem` : Température annuelle moyenne en degrés Farenheit
- `usi` : Nombre d'entreprises de plus de 20 personnes

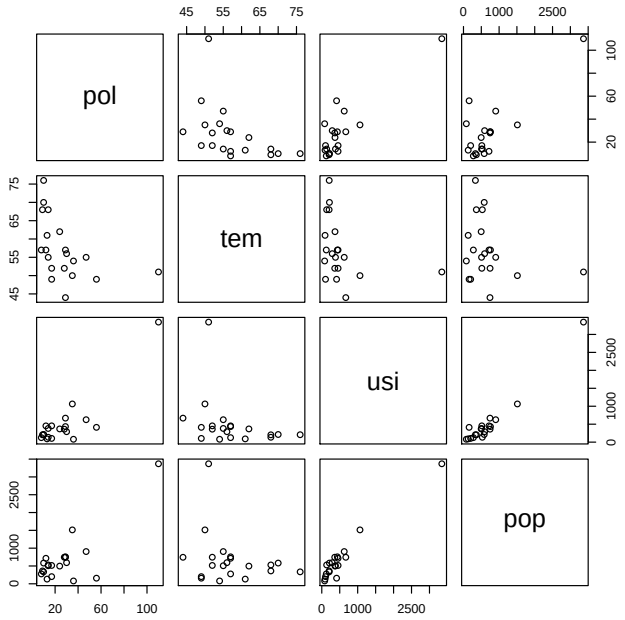
– `pop` : Population en milliers d'habitants (1970)

Retrouver ces données dans :

`http://pbil.univ-lyon1.fr/R/donnees/pollu.txt`.

Donner un commentaire à partir de la figure suivante.

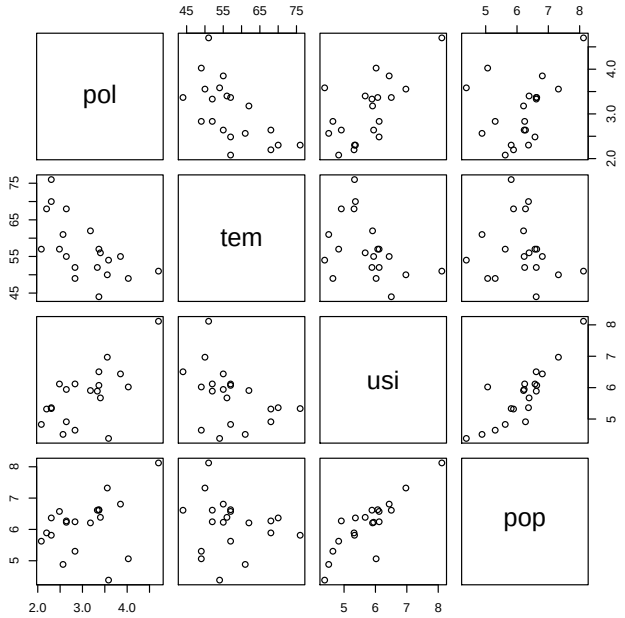
```
pairs(pollu)
```




On décide de travailler avec :

```
pol <- log(pollu$pol)
usi <- log(pollu$usi)
pop <- log(pollu$pop)
tem <- pollu$tem
pairs(cbind.data.frame(pol, tem, usi, pop))
```

Justifier ces choix et commenter le résultat.




```
lmpoltup <- lm(pol ~ tem + usi + pop)
lmpoltpu <- lm(pol ~ tem + pop + usi)
lmpolupt <- lm(pol ~ usi + pop + tem)
lmpolutp <- lm(pol ~ usi + tem + pop)
lmpolput <- lm(pol ~ pop + usi + tem)
lmpolptu <- lm(pol ~ pop + tem + usi)
cha <- c("tup", "tpu", "upt", "utp", "put", "ptu")
```

Voilà un problème sérieux : dans quel ordre doit-on introduire les explicatives dans une régression multiple ? Pour faire des comparaisons noter l'information fondamentale sur les fonctions `assign` et `get` :

`assign("a",get("b")) fait a<-b`

Les coefficients d'un modèle linéaire dépendent-ils de l'ordre d'introduction des variables ?

```
lapply(cha, function(x) get(paste("lmpol", x, sep = ""))$coefficients)
```

Les niveaux de signification de l'ANOVA dépendent-ils de l'ordre d'introduction des variables ?

```
lapply(cha, function(x) anova(get(paste("lmpol", x, sep = ""))))
```

Les tests sur les coefficients de régression dépendent-ils de l'ordre d'introduction des variables ?

```
lapply(cha, function(x) summary(get(paste("lmpol", x, sep = "")))$coefficients)
```

Comparer les corrélations entre les explicatives, les corrélations entre explicatives et variable à prédire et les coefficients de régression du modèle. Identifier la contradiction.

```
cor(cbind.data.frame(tem, usi, pop))
cor(pol, (cbind.data.frame(tem, usi, pop)))
```

Éliminer la variable perturbatrice et vérifier que le modèle à deux variables qui en résulte est sain. Tous ses indicateurs sont cohérents. Vérifier enfin que la prédiction avec deux variables est sensiblement celle du modèle complet. Un modèle statistique a pour ennemi le sur-paramétrage. Il existe des méthodes qui font ce travail de sélection automatiquement.

```
step(lm(pol ~ pop + tem + usi), k = log(20))
```

```
Start:  AIC=-23.09
pol ~ pop + tem + usi
       Df Sum of Sq      RSS      AIC
- pop   1    0.4314   3.8937 -23.7403
<none>                3.4623 -23.0932
- tem   1    0.5629   4.0252 -23.0761
- usi   1    1.4182   4.8805 -19.2226

Step:  AIC=-23.74
pol ~ tem + usi

       Df Sum of Sq      RSS      AIC
<none>                3.8937 -23.7403
- tem   1    1.3275   5.2212 -20.8686
- usi   1    1.7188   5.6124 -19.4234

Call:
lm(formula = pol ~ tem + usi)

Coefficients:
(Intercept)          tem          usi
    2.98306     -0.03556      0.36690
```

# 7 Analyse de covariance

L'exemple pédagogique est de J.D. Lebreton. On trouvera un exemple biologique dans la fiche `tdr334`.

```
method <- as.factor(rep(c("M1", "M2", "M3"), c(5, 5, 5)))
x <- c(2, 4, 5, 8, 6, 14, 16, 15, 19, 11, 20, 18, 23, 25, 24)
y <- c(5, 8, 7, 9, 11, 7, 8, 10, 13, 12, 20, 22, 26, 28, 24)
covjdl <- cbind.data.frame(x, y, method)
plot(x, y, type = "n")
text(x, y, method)
text(5, 25, "A", cex = 4)
abline(h = mean(y), lwd = 2, lty = 2, col = "red")
```

On a 15 élèves et on suppose que `x` est le niveau de départ, `y` est le niveau d'arrivée, et `m` est la méthode d'enseignement utilisée. **A** est le modèle nul.

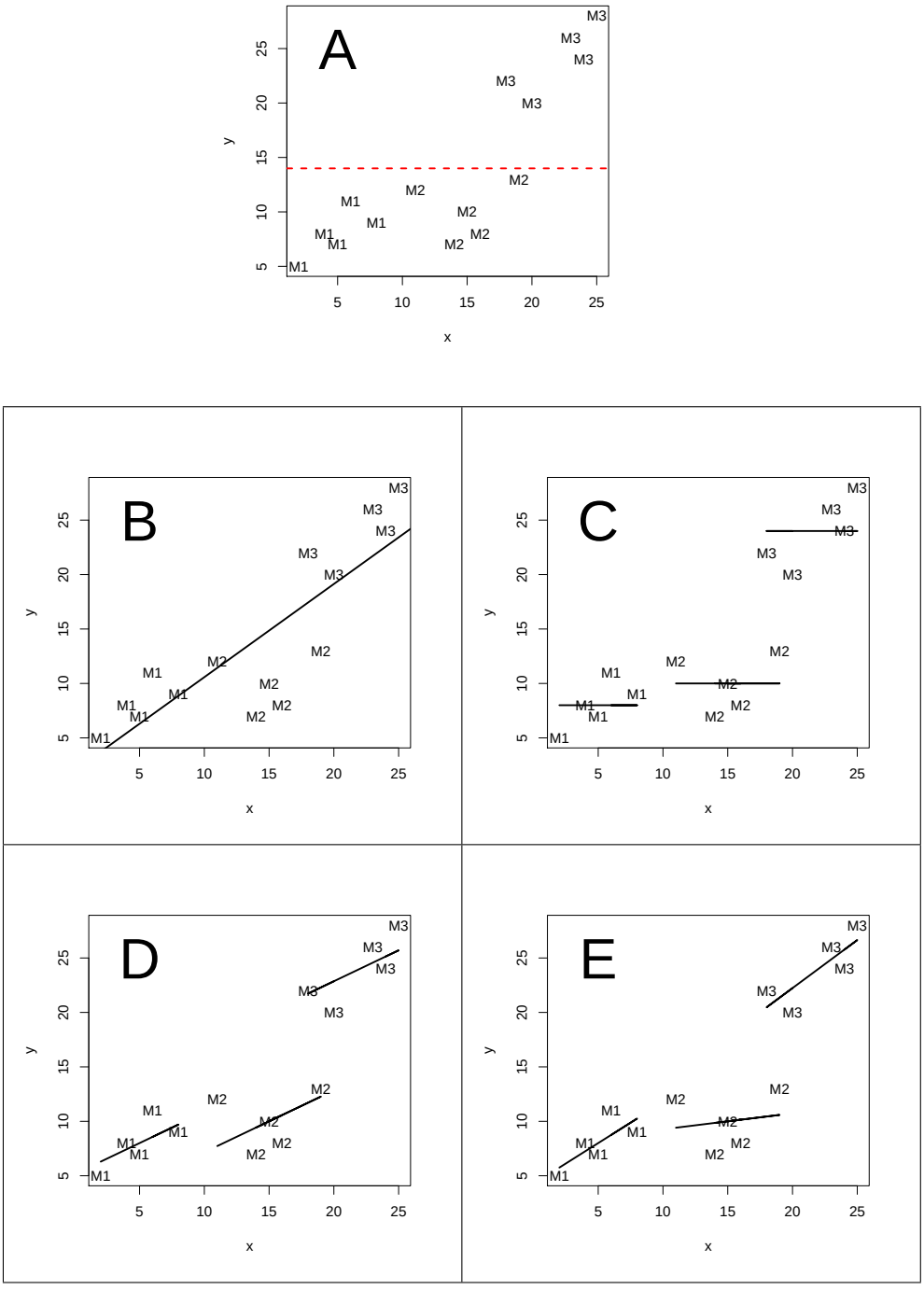



FIG. 1 – Les quatre modèles en jeu dans une analyse de covariance

Pour chacune des fenêtres de la figure 1 refaire le graphique. On peut utiliser une anova entre deux modèles linéaires. Comparer par exemple les deux

résultats :

```
anova(lm(y ~ 1), lm(y ~ x))
anova(lm(y ~ x))
```

Tester alors **A** contre **B**, **A** contre **C**, **B** contre **D**, **C** contre **D** et **D** contre **E**.

**Ce qu'on a pas le droit de faire :**

```
anova(lm(y ~ x), lm(y ~ method))
```

Expliquer pourquoi. La modélisation linéaire a encore bien d'autres ressources. Les fiches `tdr331` à `tdr3334` explorent quelques développements plus subtils.

# Références


[1] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical methods for data analysis*. Duxbury Press, Boston, 1983.

[2] W.S. Cleveland. *Visualizing data*. Hobart Press, Summit, New Jersey, 1993.

[3] W.S. Cleveland. *The elements of graphing data*. Hobart Press, Summit, New Jersey, 1994.

[4] R.D. Cook and S. Weisberg. *Residuals and influence in regression*. Chapman and Hall, New York, 1982.

[5] P. Dagnelie. *Théorie et méthodes statistiques. Exercices*. Les Presses Agronomiques de Gembloux, Gembloux, Belgique, 1981.

[6] S.H.C. du Toit, A.G.W. Steyn, and R.H. Stumpf. *Graphical Exploratory data analysis*. Springer-Verlag, New York, 1986.

[7] R. Tomassone, S. Audrain, E. Lesquoy de Turckheim, and C. Millier. *La régression*. Masson, Paris, 1992.

[8] R. Tomassone, C. Dervin, and J.P. Masson. *Biométrie Modélisation de phénomènes biologiques*. Masson, Paris, 1993.

[9] J.W. Tukey. *Exploratory data analysis*. Addison-Wesley Publishing Company, Reading, Massachussets, 1977.

[10] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55 :1–17, 1968.