

Feuille de TD : Le modèle linéaire

**Exercice 1.**

On reprend dans cet exercice le jeu de donnée *wine* utilisé dans les tutoriels R. Ce jeu de données comprend des mesures physico-chimiques réalisées sur un échantillon de 600 vins (rouges et blancs) du Portugal. Ces mesures sont complétées par une évaluation sensorielle de la qualité par un ensemble d'experts. On s'intéresse ici au lien éventuel entre la densité ( $Y$ ) et la teneur en alcool ( $z$ ). Pour cela, on considère le modèle de régression linéaire simple suivant :

$$\begin{cases} Y_i = a + bz_i + \varepsilon_i \\ \varepsilon_1, \dots, \varepsilon_n \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

1. Ecrivez ce modèle sous forme matricielle  $Y = X\theta + \varepsilon$
2. Déterminez les estimateurs  $\hat{a}$ ,  $\hat{b}$  et  $\hat{\sigma}^2$  de  $a$ ,  $b$  et  $\sigma^2$  respectivement par la méthode des moindres carrés.
3. Déterminez la loi de  $\hat{a}$ ,  $\hat{b}$  et  $\hat{\sigma}^2$ .
4. Définissez les valeurs ajustées par ce modèle et précisez leur loi.
5. Déterminez les estimateurs de  $a$ ,  $b$  et  $\sigma^2$  par la méthode du maximum de vraisemblance. Comparez ces derniers à  $\hat{a}$ ,  $\hat{b}$  et  $\hat{\sigma}^2$  obtenus précédemment.
6. Montrez que le vecteur des résidus n'est pas corrélé avec la variable explicative.
7. Testez la nullité du paramètre  $a$  au niveau 5%.
8. Construisez le test correspondant à la dernière ligne dans la sortie ci-dessous.
9. Pour un nouveau vin on a mesuré une teneur en alcool de 10.5. Construisez un intervalle de prédiction pour sa densité.

```
> summary(lm(Densite~Alcool,data=wine))

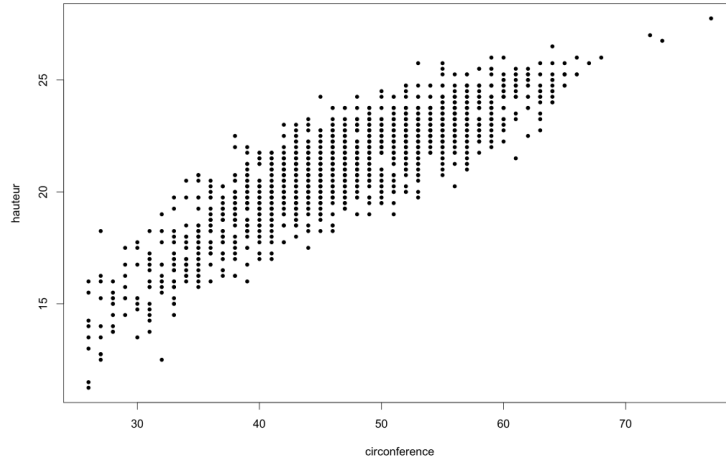
Call:
lm(formula = Densite ~ Alcool, data = wine)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0056436 -0.0015576 -0.0000468  0.0013700  0.0114534

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.013e+00  8.098e-04 1250.58  <2e-16 ***
Alcool       -1.723e-03  7.677e-05  -22.45  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002156 on 598 degrees of freedom
Multiple R-squared:  0.4573,    Adjusted R-squared:  0.4564
F-statistic: 503.9 on 1 and 598 DF,  p-value: < 2.2e-16
```

**Exercice 2.** On souhaite expliquer la hauteur  $H$  (en mètre) d'un arbre en fonction de sa circonférence (en centimètre, à 1m30 du sol) et en fonction de la racine carrée de la circonférence. On a relevé  $n = 1429$  couples  $(x_i, h_i)$ , le nuage de points étant représenté ci-dessous.



Les sorties obtenues avec le logiciel R pour cette étude sont disponibles à la fin de l'exercice. On considère le modèle de régression multiple suivant :

$$H_i = a_1 + a_2 x_i + a_3 \sqrt{x_i} + \varepsilon_i, \quad i = 1, \dots, n$$

où les  $\varepsilon_i$  sont supposés indépendants et de même loi  $\mathcal{N}(0, \sigma^2)$ .

1. Spécifiez  $H$ ,  $X$ ,  $\theta$  et  $\varepsilon$  afin d'écrire ce modèle sous la forme  $H = X\theta + \varepsilon$ .
2. On a observé

$$X'X = \begin{pmatrix} ? & ? & 9792 \\ ? & 3306476 & ? \\ ? & 471238 & 67660 \end{pmatrix}, \quad X'H = \begin{pmatrix} 30312 \\ 1461695 \\ 209685 \end{pmatrix} \text{ et } H'H = 651858.$$

- (a) Déterminez les "?" dans la matrice  $X'X$ .  
On admettra dans la suite que  $X'X$  est inversible
- (b) Que vaut la circonférence moyenne  $\bar{x}_n$  et la hauteur moyenne empirique  $\bar{h}_n$ .
3. Déterminez l'estimateur des moindres carrés  $\hat{\theta}$  pour  $\theta$  et précisez sa loi.
4. Déterminez un estimateur  $\hat{\sigma}^2$  pour  $\sigma^2$  et précisez sa loi.
5. Construisez un intervalle de confiance au niveau de confiance de 95% pour le paramètre  $a_3$ . Que pouvez-vous en conclure sur le paramètre  $a_3$ ?
6. Testez l'hypothèse  $\mathcal{H}_0 : a_2 = 0$  contre  $\mathcal{H}_1 : a_2 \neq 0$  au risque 5% de deux façons différentes.
7. Construisez un intervalle de prédiction à 95% de la hauteur  $H_0$  d'un nouvel arbre de circonférence  $x_0 = 49\text{cm}$ .

```

>mod1<-lm(ht~circ + sqrt(circ),data=arbre)

Call:
lm(formula = ht ~ circ + sqrt(circ), data = arbre)

Residuals:
    Min       1Q   Median       3Q      Max
-4.1881 -0.6881  0.0427  0.7927  3.7481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.35200    2.61444   -9.314  <2e-16 ***
circ         -0.48295    0.05793   -8.336  <2e-16 ***
sqrt(circ)    9.98689    0.78033   12.798  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.136 on 1426 degrees of freedom
Multiple R-squared:  0.7922,    Adjusted R-squared:  0.7919
F-statistic: 2718 on 2 and 1426 DF,  p-value: < 2.2e-16

> confint(mod1)
                2.5 %    97.5 %
(Intercept) -29.4805676 -19.223439
circ        -0.5965919  -0.369299
sqrt(circ)   8.4561795  11.517597

> mod0<-lm(ht~sqrt(circ),data=arbre)
> anova(mod0,mod1)
Analysis of Variance Table

Model 1: ht ~ sqrt(circ)
Model 2: ht ~ circ + sqrt(circ)
    Res.Df  RSS Df Sum of Sq  F    Pr(>F)
1    1427 1930.3
2    1426 1840.7    1    89.696 69.489 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> predict(mod1,new=data.frame(circ=49),interval="conf")
      fit      lwr      upr
1 21.89189 21.82025 21.96352

> predict(mod1,new=data.frame(circ=49),interval="pred")
      fit      lwr      upr
1 21.89189 19.66208 24.12169

```

**Exercice 3.** On considère dans cet exercice les données *Prostate* disponible dans la librairie *lasso2* de R. Ces données proviennent d'une étude qui a examiné le lien entre le niveau d'antigène prostatique spécifique et un certain nombre de mesures cliniques chez 97 hommes qui étaient sur le point de subir une prostatectomie radicale.

- lcavol : log(cancer volume)
- lweight: log(prostate weight)
- age: age
- lbph: log(benign prostatic hyperplasia amount)
- svi: seminal vesicle invasion
- lcp: log(capsular penetration)
- gleason: Gleason score
- pgg45: percentage Gleason scores 4 or 5
- lpsa: log(prostate specific antigen)

```

summary(Prostate)
      lcavol      lweight      age      lbph      svi
Min.   :-1.3471  Min.   :2.375  Min.   :41.00  Min.   :-1.3863  Min.   :0.0000
1st Qu.: 0.5128  1st Qu.:3.376  1st Qu.:60.00  1st Qu.: -1.3863  1st Qu.:0.0000
Median : 1.4469  Median :3.623  Median :65.00  Median : 0.3001  Median :0.0000
Mean   : 1.3500  Mean   :3.653  Mean   :63.87  Mean   : 0.1004  Mean   :0.2165
3rd Qu.: 2.1270  3rd Qu.:3.878  3rd Qu.:68.00  3rd Qu.: 1.5581  3rd Qu.:0.0000
Max.   : 3.8210  Max.   :6.108  Max.   :79.00  Max.   : 2.3263  Max.   :1.0000

      lcp      gleason      pgg45      lpsa
Min.   :-1.3863  Min.   :6.000  Min.   : 0.00  Min.   :-0.4308
1st Qu.: -1.3863  1st Qu.:6.000  1st Qu.: 0.00  1st Qu.: 1.7317
Median : -0.7985  Median :7.000  Median :15.00  Median : 2.5915
Mean   : -0.1794  Mean   :6.753  Mean   :24.38  Mean   : 2.4784
3rd Qu.: 1.1787  3rd Qu.:7.000  3rd Qu.:40.00  3rd Qu.: 3.0564
Max.   : 2.9042  Max.   :9.000  Max.   :100.00  Max.   : 5.5829

```

On souhaite expliquer *lpsa* en fonction des autres variables.

1. Proposez un modèle linéaire pour répondre au problème.  
Ecrivez-le sous forme matricielle.
2. Estimez les paramètres inconnus de ce modèle et précisez leur loi.
3. Définissez les valeurs ajustées par ce modèle et précisez leur loi.
4. Définissez les résidus pour ce modèle. Comment peut-on les utiliser pour vérifier les hypothèses du modèle ?
5. Construisez le test dont la p-valeur vaut 0.24964 dans `summary(modelcomplet)`
6. Que représente la Figure 1 ? Décrivez la procédure. Qu'en concluez-vous ?

7. Construisez un test de sous-modèle pour valider la conclusion de la question précédente.  
On obtient sous R la sortie suivante, qu'en concluez-vous ?

```
modcomplet <- lm(lpsa~.,data=Prostate)
summary(modcomplet)

Call:
lm(formula = lpsa ~ ., data = Prostate)

Residuals:
    Min       1Q   Median       3Q      Max
-1.73316 -0.37133 -0.01702  0.41414  1.63811

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669399   1.296381   0.516  0.60690
lcavol       0.587023   0.087920   6.677 2.11e-09 ***
lweight      0.454461   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766156   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045136   0.157464   0.287  0.77506
pgg45        0.004525   0.004421   1.024  0.30885
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared:  0.6548,    Adjusted R-squared:  0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

#### Analysis of Variance Table

```
Model 1: lpsa ~ lcavol + lweight + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
pgg45
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     93 47.785
2     88 44.163   5    3.6218 1.4434 0.2167
```

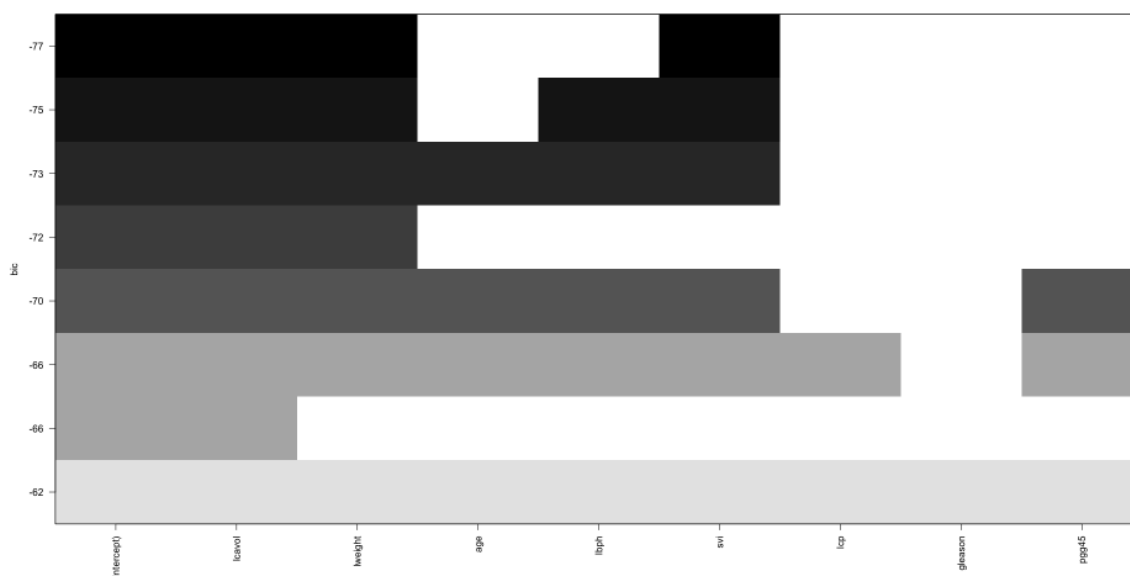


Figure 1: Figure pour exercice 3

**Exercice 4.** Sur 21 parcelles de même aire, on mesure les rendements de trois variétés différentes de blé. Les résultats sont rassemblés dans le tableau suivant :

variété 1	3,4	14	18,3	15,5	26,9	12,4	20,8	20
variété 2	5,1	22	17	21,6	30,4	17,4		
variété 3	40,1	39,6	31,6	32,2	26,4	27,0	27,8	

1. Proposez un modèle linéaire régulier  $M_1$  pour expliquer les rendements en fonction de la variété.
2. Estimez les paramètres du modèle  $M_1$  et précisez leur loi.
3. Testez s'il existe un effet variété sur le rendement.
4. On considère le modèle  $M_2$  suivant :

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, & i = 1, 2, 3, \quad j = 1, \dots, n_i \\ \varepsilon_{ij} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

où  $Y_{ij}$  est le rendement de la  $j$ ème parcelle contenant la variété  $i$ .

- (a) Que représentent les paramètres  $\mu$  et  $\alpha_i$ ?
  - (b) Ecrivez la fonction des moindres carrés à minimiser pour estimer les paramètres du modèle  $M_2$ , en fonction des  $Y_{ij}$ .
  - (c) Le modèle étant singulier, déterminez les contraintes rendant le modèle orthogonal.
  - (d) Explicitez les estimations obtenues dans la sortie `summary(lm(rendement~variete,data=A))`.
5. Testez l'hypothèse selon laquelle les rendements des variétés 2 et 3 sont identiques.

```
> summary(lm(rendement~variete-1,data=A))

Call:
lm(formula = rendement ~ variete - 1, data = A)

Residuals:
    Min       1Q   Median       3Q      Max
-13.817  -4.013  -0.500   3.587  11.483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
variete1     16.412      2.473   6.637 3.14e-06 ***
variete2     18.917      2.856   6.624 3.22e-06 ***
variete3     32.100      2.644  12.142 4.18e-10 ***
---
Residual standard error: 6.995 on 18 degrees of freedom
Multiple R-squared:  0.929, Adjusted R-squared:  0.9171
F-statistic: 78.45 on 3 and 18 DF, p-value: 1.573e-10
```

```
> summary(lm(rendement~variete,data=A))

Call:
lm(formula = rendement ~ variete, data = A)

Residuals:
    Min       1Q   Median       3Q      Max
-13.817  -4.013  -0.500   3.587  11.483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    16.412      2.473   6.637 3.14e-06 ***
variete2         2.504      3.778   0.663  0.5158
variete3        15.688      3.620   4.333  0.0004 ***
---
Residual standard error: 6.995 on 18 degrees of freedom
Multiple R-squared:  0.5362, Adjusted R-squared:  0.4847
F-statistic: 10.41 on 2 and 18 DF, p-value: 0.000993
```

### Exercice 5.

On souhaite analyser l'influence du temps et de trois espèces ligneuses d'arbre sur la décomposition de la masse d'une litière de feuilles de lierre. Pour ce faire, 24 sachets d'une masse identique de feuilles de lierre ont été constitués. Une série de 8 sachets a été déposée sous un chêne, un peuplier et un frêne respectivement. Après 2, 7, 10 et 16 semaines respectivement, deux sachets sont prélevés au hasard sous chaque arbre et la masse résiduelle est déterminée pour chacun d'eux.

Notations :  $S = \{2, 7, 10, 16\}$ ,  $A = \{"chene", "peuplier", "frene"\}$  et  $Y_{ask}$  est la masse du  $k$ -ème sachet prélevé en semaine  $s \in S$  sur l'arbre d'espèce  $a \in A$ .

1. Précisez, pour chacune des commandes ci-dessous, quel modèle est ajusté et son nom :

```
mod1 = lm(masse ~ semaine,data=Data)
mod2 = lm(masse ~ arbre + semaine,data=Data)
mod3 = lm(masse ~ semaine * arbre,data=Data)
mod4 = lm(masse ~ 1,data=Data)
mod5 = lm(masse ~ arbre, data=Data)
```

2. Quelles sont les contraintes sur les paramètres qui rendent ce modèle `mod3` orthogonal ?
3. Que représente la figure Fig.2 ? Commentez.
4. Construisez le test dont la p-valeur vaut 0.4206 dans la sortie de `anova(mod2,mod3)`.
5. A partir des différentes comparaisons de modèles proposées, déterminez le modèle qui explique le mieux les données.

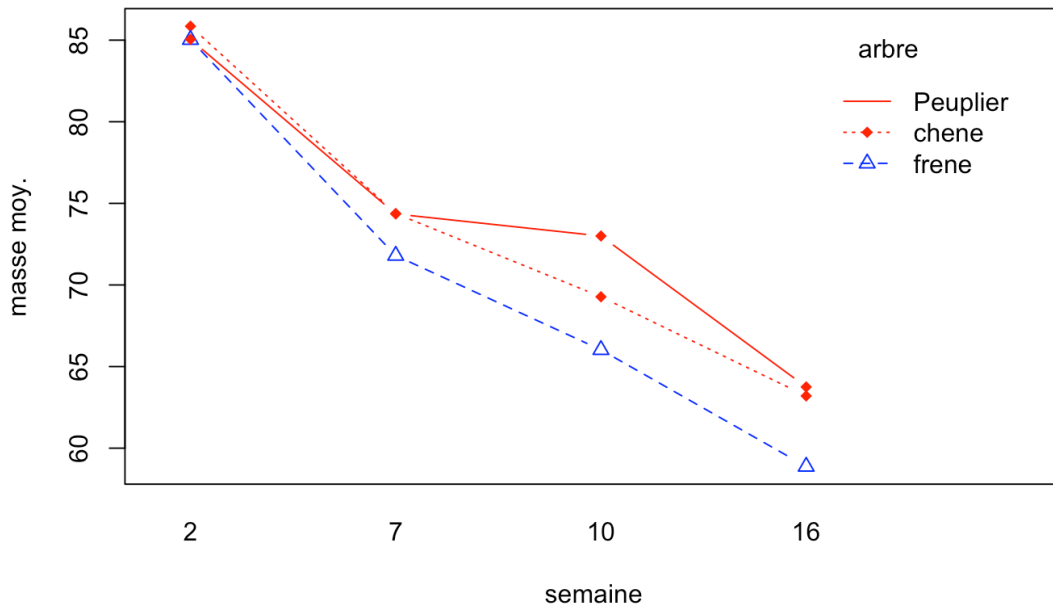


Figure 2:

#### Analysis of Variance Table

Model 1: masse ~ 1

Model 2: masse ~ semaine \* arbre

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	1857.78				
2	12	55.24	11	1802.5	35.6	1.942e-07 ***

#### Analysis of Variance Table

Model 1: masse ~ semaine

Model 2: masse ~ arbre + semaine

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	20	142.115				
2	18	85.376	2	56.739	5.9812	0.01019 *

#### Analysis of Variance Table

Model 1: masse ~ semaine

Model 2: masse ~ semaine \* arbre

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	20	142.115				
2	12	55.236	8	86.878	2.3593	0.08726 .

#### Analysis of Variance Table

Model 1: masse ~ 1

Model 2: masse ~ arbre

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	1857.8				
2	21	1801.0	2	56.739	0.3308	0.722

#### Analysis of Variance Table

Model 1: masse ~ arbre + semaine

Model 2: masse ~ semaine \* arbre

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	85.376				
2	12	55.236	6	30.139	1.0913	0.4206

#### Analysis of Variance Table

Model 1: masse ~ 1

Model 2: masse ~ arbre + semaine

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	1857.78				
2	18	85.38	5	1772.4	74.736	2.117e-11 ***

#### Analysis of Variance Table

Model 1: masse ~ arbre

Model 2: masse ~ arbre + semaine

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	1801.04				
2	18	85.38	3	1715.7	120.57	4.167e-12 ***

#### Analysis of Variance Table

Model 1: masse ~ 1

Model 2: masse ~ semaine

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	1857.78				
2	20	142.11	3	1715.7	80.483	2.449e-11 ***

**Exercice 6.** On considère un modèle d'analyse de covariance défini par

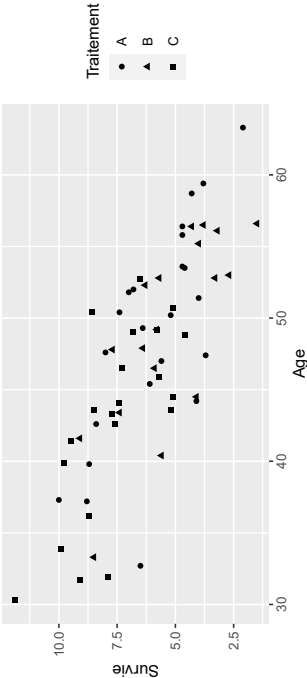
$$Y_{ij} = a_i z_{ij} + b_i + \varepsilon_{ij}, \quad i \in \{1, 2\}, \quad j \in \{1, \dots, n_i\}.$$

1. Écrivez ce modèle sous forme matricielle.
2. Commentez les différentes composantes de ce modèle: que représentent les coefficients  $a_i$ ,  $b_i$  ?
3. Proposez un estimateur du vecteur  $(a_1, b_1, a_2, b_2)'$ .
4. Explicitez un test pour l'hypothèse  $H_0 : a_1 = a_2$  et  $b_1 = b_2$ .  
Que devient le modèle d'analyse de covariance sous  $H_0$  ?

**Exercice 7.** On étudie la durée de survie  $Y$  de femmes atteintes de cancer du sein soumises à trois traitements, A, B et C. On a aussi l'information sur l'âge  $x$  d'apparition d'un cancer.

1. Sans tenir compte de l'âge d'apparition du cancer, testez l'existence d'un effet traitement.
2. On soupçonne un lien entre l'âge d'apparition et la durée de survie. On propose d'étudier plusieurs modèles dont les sorties sous R sont reportés en fin d'exercice (Modèle1, ..., Modèle5).  
Ecrivez le modèle considéré dans chacun des cas et donnez une interprétation.
3. Au vu des différents résultats, quel modèle retient-on pour cette étude?

Sorties pour l'exercice Cancer du sein



Modele 2

```
mod2 = lm(Survie~Age,data=CancerSein)
summary(mod2)

## Call:
## lm(formula = Survie ~ Age, data = CancerSein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9438 -0.7789  0.1779  0.9469  3.0697
##
## Coefficients:
##      (Intercept)  Estimate Std. Error t value Pr(>|t|)
##      Age          -0.22110    0.02414  -9.159 5.34e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.422 on 60 degrees of freedom
## Multiple R-squared:  0.583, Adjusted R-squared:  0.576
## F-statistic: 83.88 on 1 and 60 DF, p-value: 5.339e-13
```

Modele 3

```
mod3 = lm(Survie~Traitement + Age + Age:Traitement, data=CancerSein)
summary(mod3)

## Call:
## lm(formula = Survie ~ Traitement + Age + Age:Traitement, data = CancerSein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6985 -0.8347  0.1017  1.0374  2.5638
##
## Coefficients:
##      (Intercept)  Estimate Std. Error t value Pr(>|t|)
##      TraitementB  1.79530    3.23055   0.556  0.581
##      TraitementC  1.09928    2.83590   0.388  0.700
##      Age          -0.18809    0.03947  -4.766 1.38e-05 ***
##      TraitementB:Age -0.04776    0.06510  -0.734  0.466
##      TraitementC:Age -0.01380    0.06164  -0.224  0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.416 on 56 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.5797
## F-statistic: 17.83 on 5 and 56 DF, p-value: 1.551e-10
```

Modele 1

```
mod1 = lm(Survie~Traitement + Age, data=CancerSein)
summary(mod1)

## Call:
## lm(formula = Survie ~ Traitement + Age, data = CancerSein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77726 -0.82660  0.07026  0.94507  2.58343
##
## Coefficients:
##      (Intercept)  Estimate Std. Error t value Pr(>|t|)
##      TraitementB -0.55235    0.44002  -1.255  0.214
##      TraitementC  0.40722    0.45074   0.903  0.370
##      Age          -0.20450    0.02583  -7.917 8.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.398 on 58 degrees of freedom
## Multiple R-squared:  0.6104, Adjusted R-squared:  0.5903
## F-statistic: 30.3 on 3 and 58 DF, p-value: 6.504e-12
```



Modele 4

```
mod4 = lm(Survie~Traitement + Age,Traitement, data=CancerSein)
summary(mod4)

## Call:
## lm(formula = Survie ~ Traitement + Age,Traitement, data = CancerSein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6985 -0.8347  0.1017  1.0374  2.5638
##
## Coefficients:
##      (Intercept)      Estimate Std. Error t value Pr(>|t|)
##      TraitementB      1.79530    3.23055   0.556  0.581
##      TraitementC      1.09928    2.83590   0.388  0.700
##      TraitementA:Age -0.18809    0.03947  -4.766 1.38e-05 ***
##      TraitementB:Age -0.23585    0.05177  -4.556 2.87e-05 ***
##      TraitementC:Age -0.20189    0.04735  -4.264 7.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 1.416 on 56 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.5797
## F-statistic: 17.83 on 5 and 56 DF,  p-value: 1.551e-10
```

Modele 5

```
mod5=lm(Survie~Traitement,data=CancerSein)
summary(mod5)

## Call:
## lm(formula = Survie ~ Traitement, data = CancerSein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.7957 -1.5704  0.0929  1.1335  4.3429
##
## Coefficients:
##      (Intercept)      Estimate Std. Error t value Pr(>|t|)
##      TraitementB      0.6012    0.6292  -0.955  0.34325
##      TraitementC      1.6615    0.6035   2.753  0.00783 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 2 on 59 degrees of freedom
## Multiple R-squared:  0.1894, Adjusted R-squared:  0.162
## F-statistic: 6.895 on 2 and 59 DF,  p-value: 0.002037
```

Comparaison des modèles

```
anova(mod2,mod3)

## Analysis of Variance Table
##
## Model 1: Survie ~ Age
## Model 2: Survie ~ Traitement + Age + Age:Traitement
## Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      60 121.36
## 2      56 112.28  4    9.0743  1.1314 0.3511
anova(mod1,mod3)

## Analysis of Variance Table
##
## Model 1: Survie ~ Traitement + Age
## Model 2: Survie ~ Traitement + Age + Age:Traitement
## Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      58 113.37
## 2      56 112.28  2    1.0879  0.2713 0.7634
anova(mod5,mod3)

## Analysis of Variance Table
##
## Model 1: Survie ~ Traitement
## Model 2: Survie ~ Traitement + Age + Age:Traitement
## Res.Df  RSS Df Sum of Sq    F    Pr(>F)
## 1      59 235.89
## 2      56 112.28  3   123.61 20.549 4.178e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(mod2,mod1)

## Analysis of Variance Table
##
## Model 1: Survie ~ Age
## Model 2: Survie ~ Traitement + Age
## Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      60 121.36
## 2      58 113.37  2    7.9863  2.0429 0.1389
anova(mod5,mod1)

## Analysis of Variance Table
##
## Model 1: Survie ~ Traitement
## Model 2: Survie ~ Traitement + Age
## Res.Df  RSS Df Sum of Sq    F    Pr(>F)
## 1      59 235.89
## 2      58 113.37  1   122.52 62.681 8.434e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(CancerSein)

##   Traitement   Age   Survie
## Length:62    Min.   30.30  Min.   1.500
## Class :character 1st Qu.:42.77 1st Qu.: 4.625
## Mode :character  Median :47.70 Median : 6.200
##              Mean  :46.99  Mean  : 6.284
##              3rd Qu.:52.60 3rd Qu.: 7.850
##              Max.   :63.30 Max.   :11.900
##
## table(CancerSein$Traitement)
##
##      A  B  C
## 23 18 21
```