Feuille de Travaux Dirigés n° 4
Analyse de la covariance

Les exercices 1 et 2 sont tirés du livre *Modèle linéaire : Comparaison de groupes et régression* de B. Prum aux Éditions de l'INSERM

Exercice IV.1. Cancer du sein

On étudie la durée de survie Y de femmes atteintes de cancer du sein soumises à trois traitements, A, B et C. Ces durées figurent dans le tableau suivant dans les colonnes *Survie* ; on a aussi indiqué l'âge X d'apparition d'un cancer.

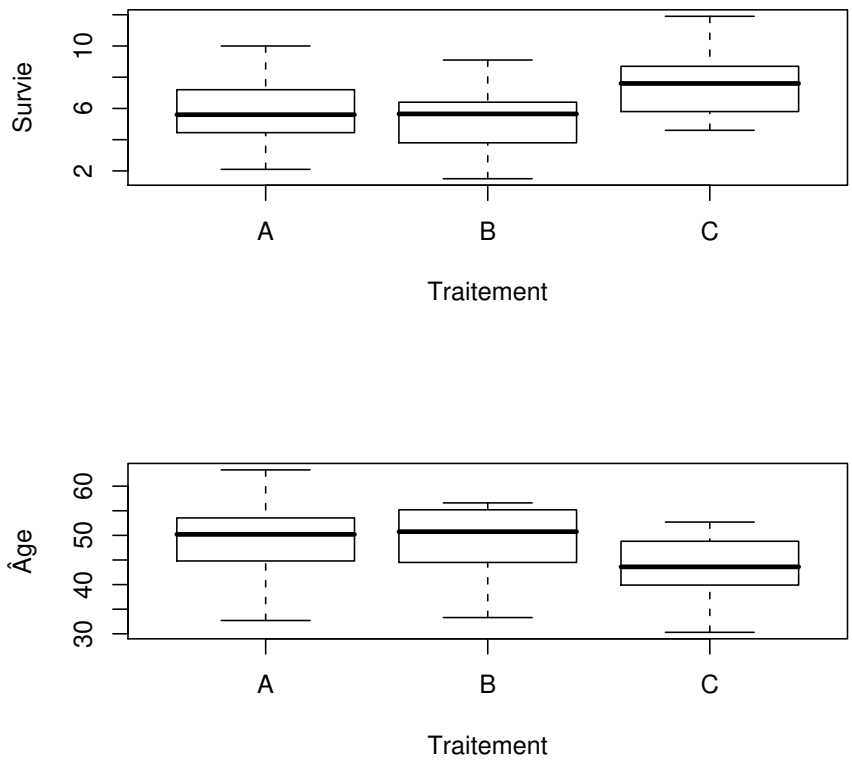
Traitement A		Traitement B		Traitement C	
Âge	Survie	Âge	Survie	Âge	Survie
32,7	6,5	33,3	8,5	30,3	11,9
37,2	8,8	40,4	5,6	31,7	9,1
37,3	10,0	41,6	9,1	31,9	7,9
39,8	8,7	43,4	7,4	33,9	9,9
42,6	8,4	44,5	4,1	36,2	8,7
44,2	4,1	46,5	5,9	39,9	9,8
45,4	6,1	47,8	7,7	41,4	9,5
47,0	5,6	47,9	6,4	42,6	7,6
47,4	3,7	49,2	5,8	43,3	7,7
47,6	8,0	52,3	6,3	43,6	5,2
49,3	6,4	52,8	5,7	43,6	8,5
50,2	5,2	52,8	3,3	44,1	7,4
50,4	7,4	53,0	2,7	44,5	5,1
51,4	4,0	55,2	4,0	45,9	5,7
51,8	7,0	56,1	3,2	46,5	7,3
52,0	6,8	56,4	4,3	48,8	4,6
53,5	4,6	56,5	3,8	49,0	6,8
53,6	4,7	56,6	1,5	49,2	5,8
55,8	4,7			50,4	8,6
56,4	4,7			50,7	5,1
58,7	4,3			52,7	6,5
59,4	3,8				
63,3	2,1				

1. Récupérer les données dans R en exécutant les instructions suivantes[1].

```
> options(contrasts = c("contr.sum", "contr.poly"))
> Chemin <- "C:\\..."
> CancerSein <- read.table(paste(Chemin, "CancerSein.CSV",
+     sep = ""), sep = ";", dec = ".", quote = "\"", header = T)
```

2. Représenter graphiquement les données à l'aide d'un nuage de points où l'on spécifiera le traitement reçu. On pourra obtenir la représentation ci-dessous. Calculer la durée moyenne de survie dans chaque groupe ? Sans tenir compte de l'âge d'apparition du cancer, tester l'existence d'un effet traitement.

```
> layout(c(1, 2))
> plot(Survie ~ Traitement, data = CancerSein)
> plot(Âge ~ Traitement, data = CancerSein)
> layout(1)
> tapply(CancerSein$Âge, CancerSein$Traitement, mean)
> aov1 <- aov(Survie ~ Traitement, data = CancerSein)
> shapiro.test(residuals(aov1))
> library(car)
> bartlett.test(residuals(aov1), CancerSein$Traitement)
```



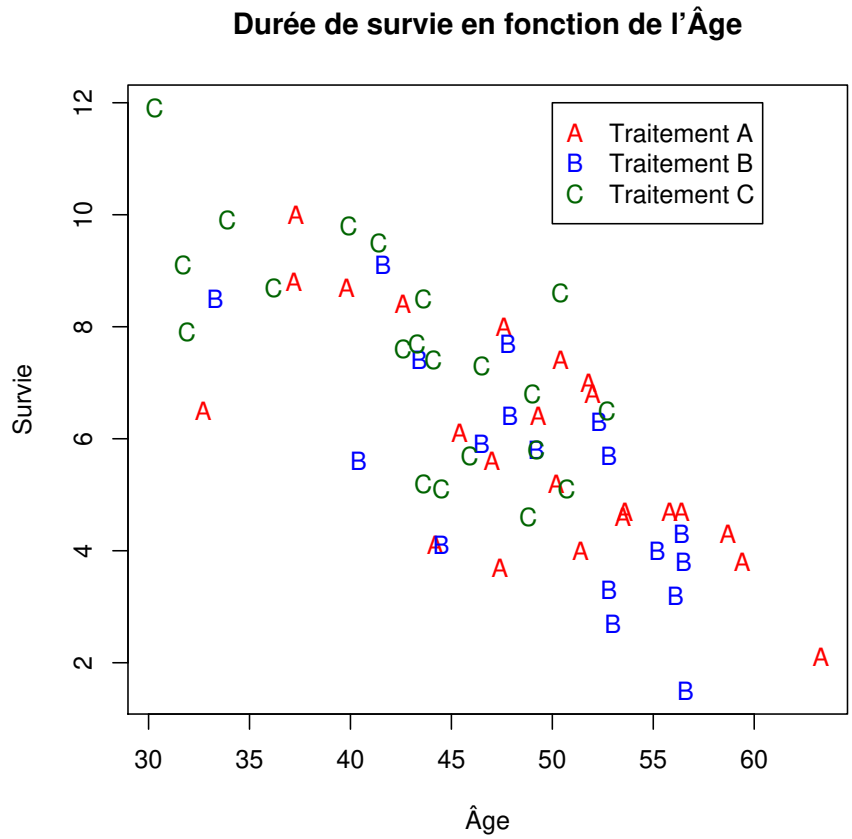
[1]Il faut remplacer `"C :\\..."` par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.

```
> summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Traitement	2	55.13	27.57	6.89	0.0020
Residuals	59	235.89	4.00		

3. Représenter graphiquement les données à l'aide d'un diagramme de la durée de survie en fonction de l'âge où l'on spécifiera de surcroît le traitement reçu.

```
> plot(Survie ~ Âge, main = "Durée de survie en fonction de l'Âge",
+     data = CancerSein, type = "n")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "A"), col = "red", pch = "A")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "B"), col = "blue", pch = "B")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "C"), col = "darkgreen", pch = "C")
> legend(50, 12, c("Traitement A", "Traitement B", "Traitement C"),
+     pch = "ABC", col = c("red", "blue", "darkgreen"),
+     cex = 1)
```



4. On soupçonne un lien entre l'âge d'apparition et la durée de survie, quels modèles peut-on envisager ? Étudier en particulier les modèles qui comportent les termes suivants :

a. Effet linéaire de l'âge.

```
> (mod1 <- lm(Survie ~ Âge, data = CancerSein))
> shapiro.test(residuals(mod1))
```

b. Effet linéaire de l'âge et effets principaux des traitements.

```
> (mod2 <- lm(Survie ~ Âge + Traitement, data = CancerSein))
> shapiro.test(residuals(mod2))
```

c. Effets linéaires de l'âge différents en fonction des traitements et effets principaux des traitements. Ce modèle permet de savoir si au sein de chacun des groupes le facteur Âge est significatif.

```
> (mod3 <- lm(Survie ~ Traitement + Âge:Traitement, data = CancerSein))
> shapiro.test(residuals(mod3))
```

d. Effets linéaires de l'âge commun pour tous les groupes plus un terme correctif pour chacun des groupes et effets principaux des traitements. Ce modèle permet de savoir si l'on peut considérer que l'intensité de la dépendance du temps de survie est associée au traitement ou simplement aux individus indépendamment des traitements qu'ils ont reçus.

```
> (mod4 <- lm(Survie ~ Âge + Traitement + Âge:Traitement,
+     data = CancerSein))
> shapiro.test(residuals(mod4))
```

5. Pour déterminer le modèle à utiliser nous utilisons le critère AIC :

```
> step(mod4)
```

Nous retenons donc le modèle du **3.b.**. Nous étudions plus précisement les résultats qui découlent de ce choix.

```
> summary(mod2)
```

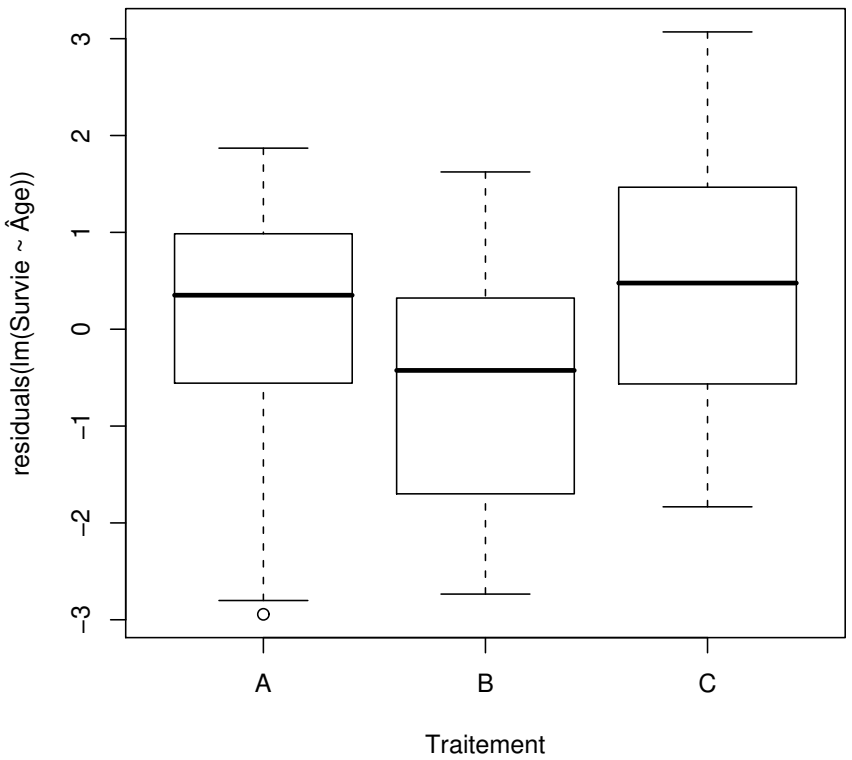
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.8678	1.2280	12.92	0.0000
Âge	-0.2045	0.0258	-7.92	0.0000
Traitement1	0.0484	0.2505	0.19	0.8475
Traitement2	-0.5040	0.2670	-1.89	0.0641

```
> anova(mod2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Âge	1	169.67	169.67	86.80	0.0000
Traitement	2	7.99	3.99	2.04	0.1389
Residuals	58	113.37	1.95		

Que dire des résultats obtenus à la question **1.**, en particulier de la significativité des effets des traitements ? On pourra représenter la boîte à moustaches des effets des traitements ajustés pour les variations de l'âge. Quelles sont les différences avec les boîtes à moustaches construites en **1.** ?

```
> plot(residuals(lm(Survie ~ Âge)) ~ Traitement, data = CancerSein)
```



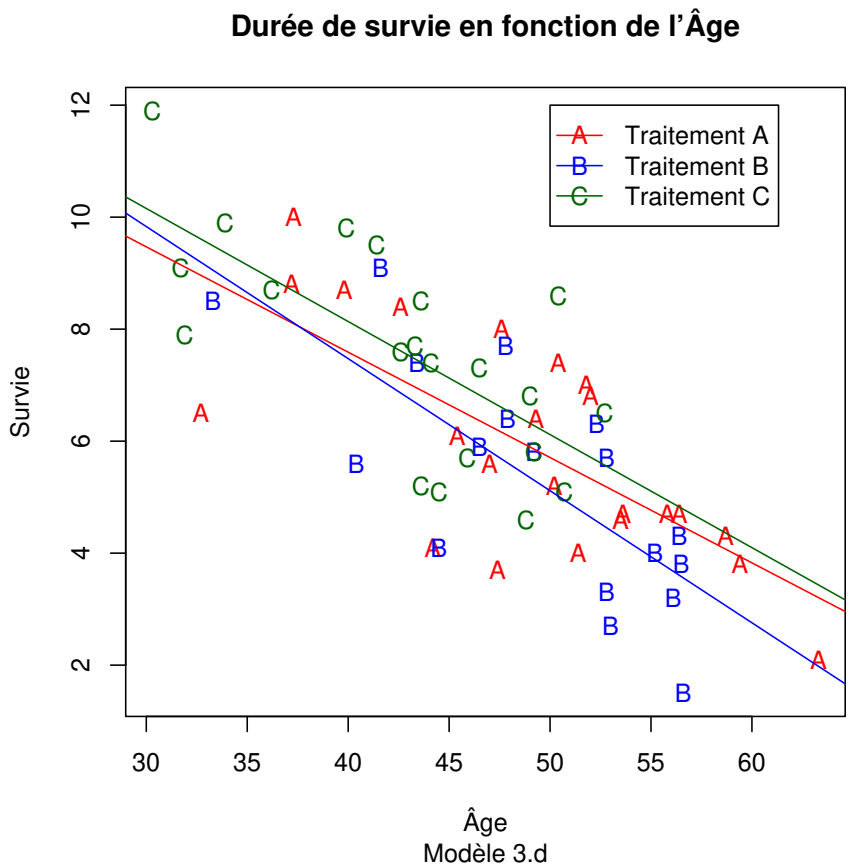
6. En utilisant le modèle du **3.d.**, déterminer s'il y a une dépendance de l'intensité de l'effet de l'âge par rapport au traitement utilisé ? On décidera de la significativité de cette dépendance à un seuil de $\alpha = 5$ % puis on construira les représentations graphiques ci-dessous à l'aide des valeurs prédites par le modèle du **3.d.** pour chacun des groupes et des observations puis à l'aide des valeurs prédites par le modèle du **3.b.** pour chacun des groupes et des observations. Interpréter ces graphiques à l'aide des résultats du test précédent.

```
> summary(mod4)

> anova(mod4)

> plot(Survie ~ Âge, main = "Durée de survie en fonction de l'Âge",
+     data = CancerSein, type = "n", sub = "Modèle 3.d")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "A"), col = "red", pch = "A")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
```

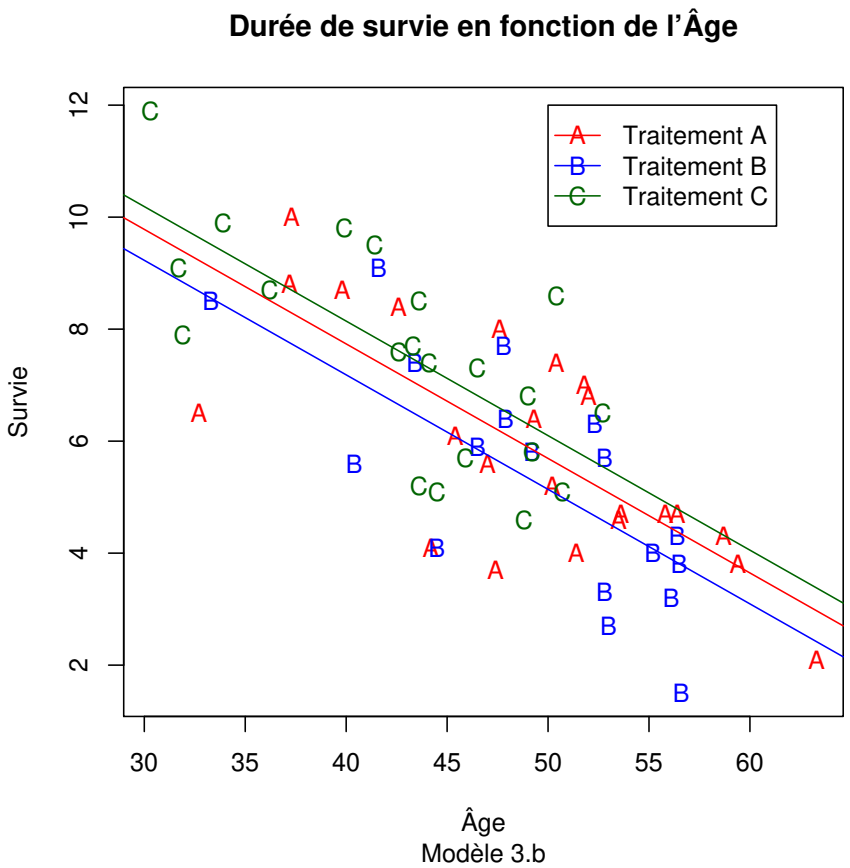
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.0771	1.2759	12.60	0.0000
Âge	-0.2086	0.0268	-7.77	0.0000
Traitement1	-0.9649	1.7040	-0.57	0.5735
Traitement2	0.8304	1.9573	0.42	0.6730
Âge :Traitement1	0.0205	0.0352	0.58	0.5623
Âge :Traitement2	-0.0272	0.0402	-0.68	0.5005

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Âge	1	169.67	169.67	84.62	0.0000
Traitement	2	7.99	3.99	1.99	0.1460
Âge :Traitement	2	1.09	0.54	0.27	0.7634
Residuals	56	112.28	2.01		

```
+     "B"), col = "blue", pch = "B")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "C"), col = "darkgreen", pch = "C")
> (mod5 <- lm(Survie ~ Traitement + Âge:Traitement - 1,
+     data = CancerSein))
> coe <- coefficients(mod5)
> abline(coe[1], coe[4], col = "red")
> abline(coe[2], coe[5], col = "blue")
> abline(coe[3], coe[6], col = "darkgreen")
> legend(50, 12, c("Traitement A", "Traitement B", "Traitement C"),
+     pch = "ABC", lty = c(1, 1, 1), col = c("red", "blue",
+         "darkgreen"), cex = 1)
```




```
> plot(Survie ~ Âge, main = "Durée de survie en fonction de l'Âge",
+     data = CancerSein, type = "n", sub = "Modèle 3.b")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "A"), col = "red", pch = "A")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "B"), col = "blue", pch = "B")
> points(Survie ~ Âge, data = subset(CancerSein, Traitement ==
+     "C"), col = "darkgreen", pch = "C")
> (mod6 <- lm(Survie ~ Traitement + Âge - 1, data = CancerSein))
> coe2 <- coefficients(mod6)
> abline(coe2[1], coe2[4], col = "red")
> abline(coe2[2], coe2[4], col = "blue")
> abline(coe2[3], coe2[4], col = "darkgreen")
> legend(50, 12, c("Traitement A", "Traitement B", "Traitement C"),
+     pch = "ABC", lty = c(1, 1, 1), col = c("red", "blue",
+         "darkgreen"), cex = 1)
```



. . . . . .

Exercice IV.2. Le SIDA du chat

On mesure le taux de leucocytes T4 chez le chat X_2 jours après avoir inoculé à l'animal le virus FeLV, analogue au HIV. On appelle Y le logarithme népérien de ce taux. Le tableau ci-dessous donne les mesures faites sur 17 chats mâles et 15 chats femelles. Le facteur *Sexe* est noté X_1.

Mâles		Femelles	
Jours	Ln(Taux de T4)	Jours	Ln(Taux de T4)
44	4,66	84	3,45
317	3,08	47	3,89
292	1,28	20	3,79
179	3,17	209	3,79
39	5,59	106	3,81
257	2,88	343	0,61
354	1,60	325	2,04
349	3,48	346	0,41
195	3,39	151	2,67
245	3,47	267	0,89
270	3,20	80	4,39
166	2,90	249	2,56
57	4,83	341	0,28
198	2,96	189	2,43
20	5,17	50	3,85
187	3,44		
270	3,18		

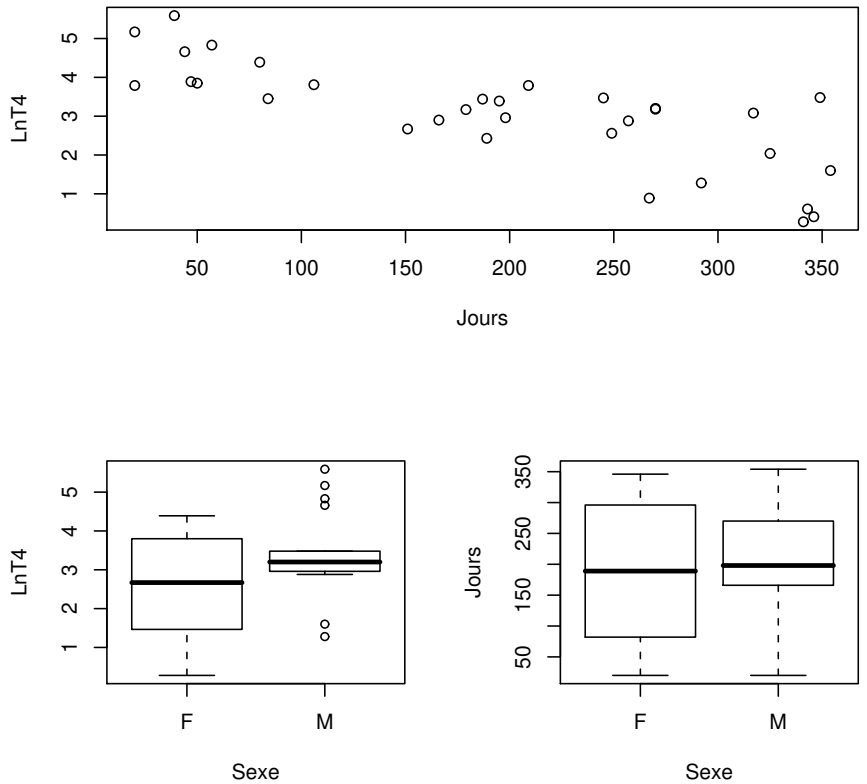
1. Récupérer les données dans R en exécutant les instructions suivantes[2].

```
> Chemin <- "C:\\..."
> options(contrasts = c("contr.sum", "contr.poly"))

> SidaChat <- read.table(paste(Chemin, "SidaChat.CSV",
+     sep = ""), sep = ";", dec = ".", quote = "\"", header = T)
```

2. Quelle est la nature de chaque facteur ? Étudier les données par rapport à chacun des facteurs.

```
> layout(matrix(c(1, 1, 2, 3), byrow = T, nrow = 2))
> plot(LnT4 ~ Jours, data = SidaChat)
> plot(LnT4 ~ Sexe, data = SidaChat)
> plot(Jours ~ Sexe, data = SidaChat)
> layout(1)
```

[2]Il faut remplacer "C :\\..." par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.



2. On s'intéresse désormais à des modèles utilisant les deux facteurs simultanément. On peut envisager chacun des cas suivants :

a. On considère une droite dont la pente ne dépend pas du sexe du chat mais dont l'ordonnée à l'origine dépend du sexe du chat.

$\boldsymbol{\alpha}$**.** $Y_{i,j} = \alpha_i + \beta_1 X_{2,i,j} + \epsilon_{i,j}$

b. On considère deux droites dont les pentes dépendent du sexe du chat mais dont l'ordonnée à l'origine ne dépend pas du sexe du chat.

$\boldsymbol{\beta}$**.** $Y_{i,j} = \beta_0 + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$

c. On considère deux droites de régression, l'une pour les chats mâles, l'autre pour les chats femelles.

$\boldsymbol{\gamma}$**.** $Y_{i,j} = \alpha_i + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$

d. On considère deux droites dont les pentes ne dépendent pas du sexe du chat mais dont l'ordonnée à l'origine dépend du sexe du chat comme un écart à une valeur moyenne commune aux deux sexes.

$\boldsymbol{\delta}$**.** $Y_{i,j} = \beta_0 + \alpha_i + \beta_1 X_{2,i,j} + \epsilon_{i,j}$.

e. On considère deux droites dont les pentes dépendent du sexe du chat comme un écart par rapport à une valeur de la pente que l'on prendrait comme commune aux deux sexes et dont l'ordonnée à l'origine dépend du sexe du chat comme un écart à une valeur moyenne commune aux deux sexes.

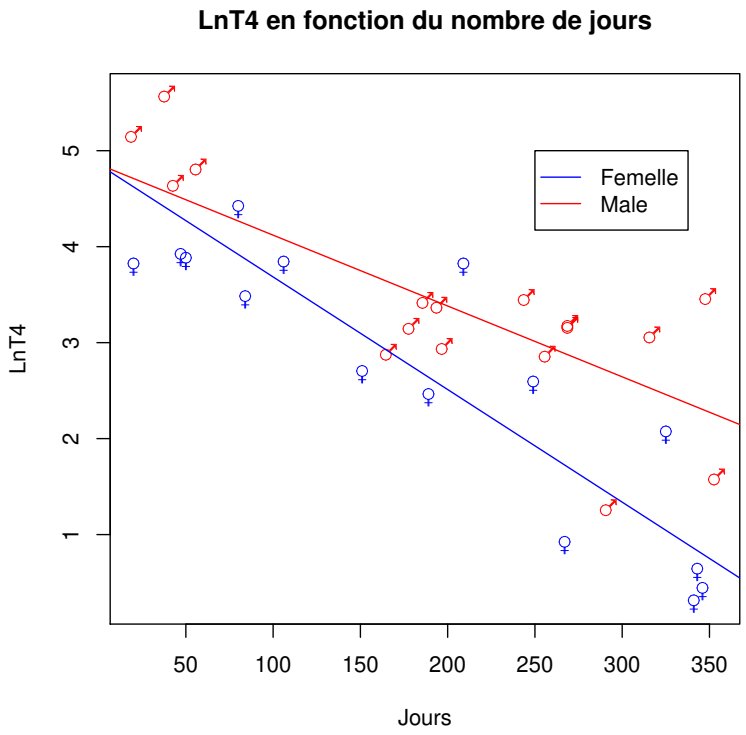
ϵ. $Y_{i,j} = \beta_0 + \alpha_i + \beta_1 X_{2,i,j} + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$

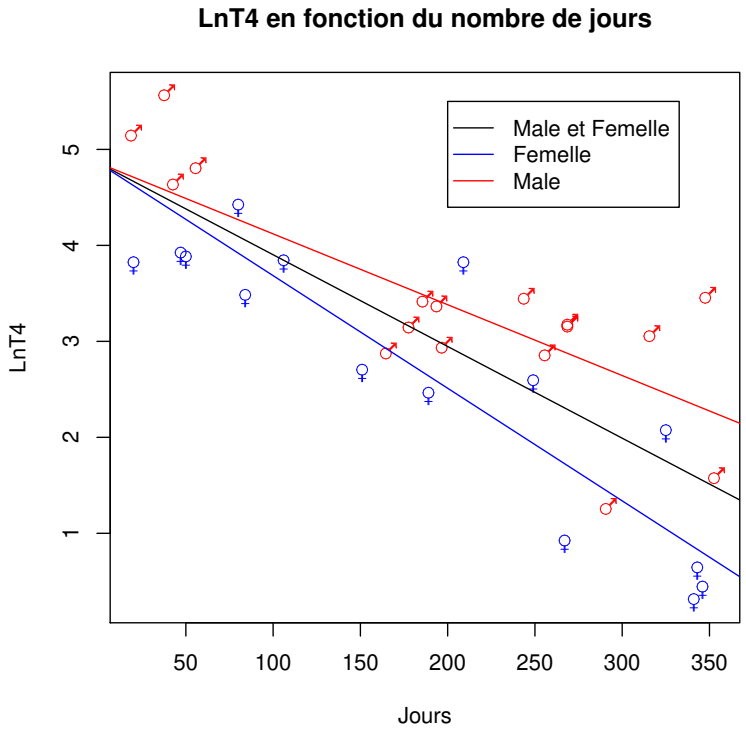
f. On considère deux droites dont les pentes dépendent du sexe du chat comme un écart par rapport à une valeur de la pente que l'on prendrait comme commune aux deux sexes et dont l'ordonnée à l'origine ne dépend pas du sexe du chat.

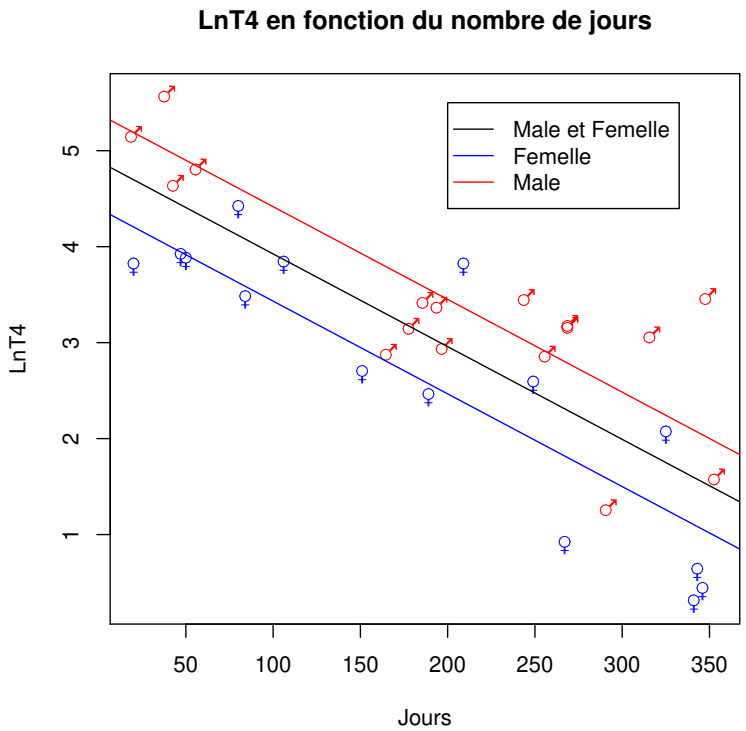
ϕ. $Y_{i,j} = \beta_0 + \beta_1 X_{2,i,j} + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$

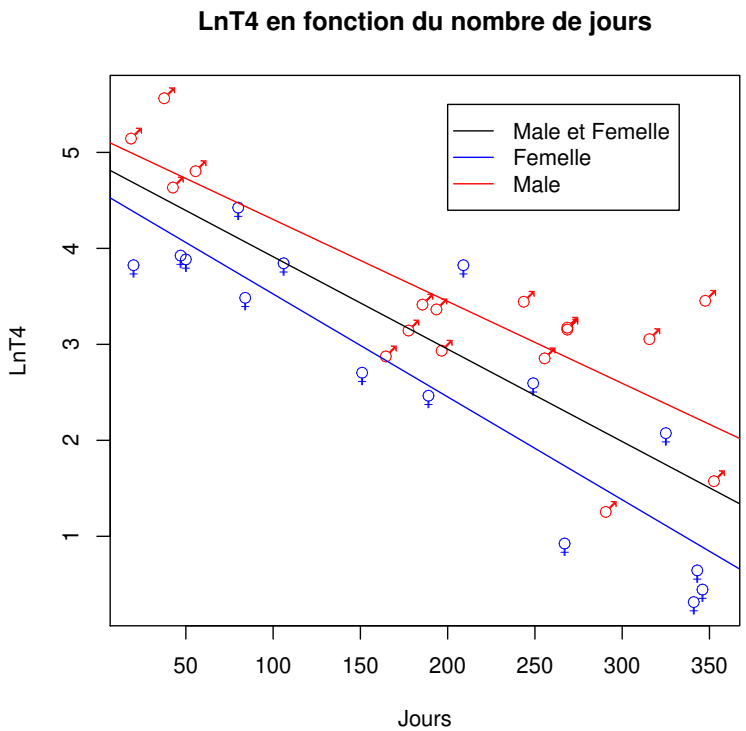
Pour alléger les notations on a noté de manière identique les paramètres dans chacun des modèles, leur valeur réelle dépendant en fait du modèle $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}$ considéré.
Chacune des équations ci-dessus est valable pour $1 \leq i \leq 2,\ 1 \leq j \leq n_i$ avec $n_1 = 17$ et $n_2 = 15$.
Associer à chacune des formes mathématiques des modèles ci-dessous son interprétation graphique parmis celles qui figurent dans la liste ci-dessous. Puis reproduire ces graphiques avec R. Vous pourrez compléter les instructions suivantes :

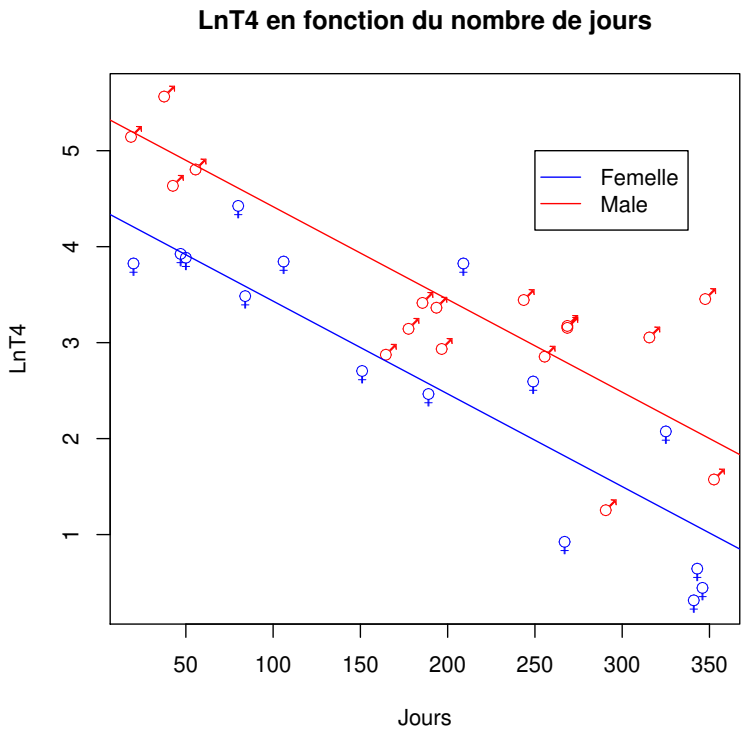
```
> plot(LnT4 ~ Jours, main = "LnT4 en fonction du nombre de jours",
+     data = SidaChat, type = "n")
> text(SidaChat$Jours, SidaChat$LnT4, ifelse(SidaChat$Sexe ==
+     "F", "\\VE", ""), col = "blue", vfont = c("serif",
+     "plain"), cex = 1.25)
> text(SidaChat$Jours, SidaChat$LnT4, ifelse(SidaChat$Sexe ==
+     "M", "\\MA", ""), col = "red", vfont = c("serif",
+     "plain"), cex = 1.25)
> legend(250, 5, c("Femelle", "Male"), pch = "", lty = c(1,
+     1), col = c("blue", "red"), cex = 1)
```

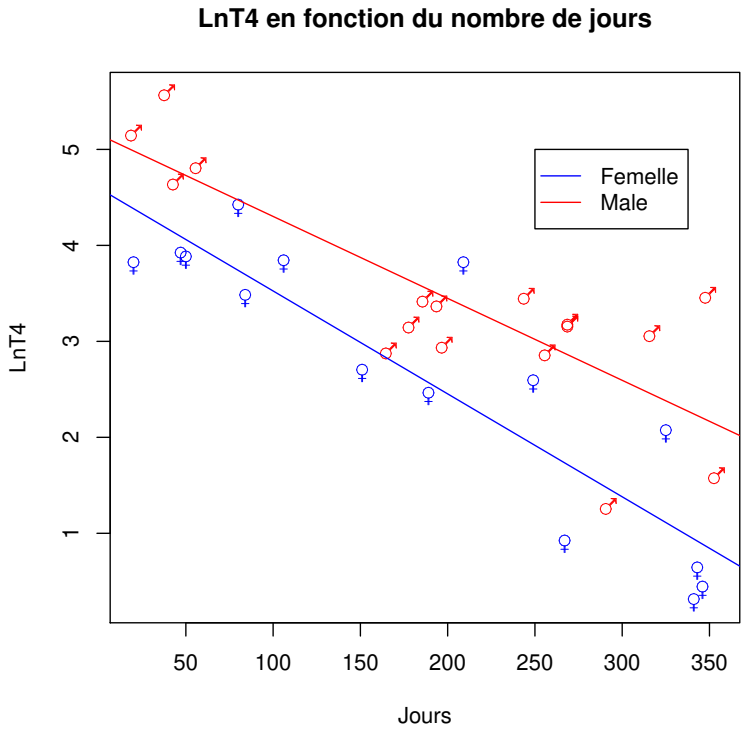

LnT4 en fonction du nombre de jours
Femelle
Male
LnT4
Jours
1
2
3
4
5
50
100
150
200
250
300
350


LnT4 en fonction du nombre de jours
Male et Femelle
Femelle
Male
LnT4
Jours
1
2
3
4
5
50
100
150
200
250
300
350


LnT4 en fonction du nombre de jours
Male et Femelle
Femelle
Male
LnT4
1
2
3
4
5
50
100
150
200
250
300
350
Jours


LnT4 en fonction du nombre de jours
Male et Femelle
Femelle
Male
LnT4
1
2
3
4
5
50
100
150
200
250
300
350
Jours





On fait les hypothèses classiques sur les erreurs : elles sont indépendantes et de même loi. Leur loi commune étant une loi normale centrée de variance σ^2, la valeur de σ^2 dépendant elle aussi du modèle $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}$ considéré.

On pensera donc à bien vérifier que les conditions d'utilisation de chacun des modèles sont bien vérifiées.

3. Déterminer quels sont les modèles pertinents qui comportent le moins de termes superflus pour procéder à l'étude du SIDA du chat.

4. On constate que le modèle $\boldsymbol{\epsilon}$**.** est un des modèles à retenir. L'évolution du SIDA du chat est-il lié au sexe du chat ? Interpréter ce résultat à l'aide de la représentation graphique associée à ce modèle.

5. Le modèle $\boldsymbol{\delta}$ peut être considéré comme un sous-modèle du modèle $\boldsymbol{\epsilon}$. En utilisant la commande `anova`, déterminer si l'on peut faire l'hypothèse que les droites peuvent avoir une ordonnée à l'origine commune. Interpréter cette hypothèse, puis obtenir la représentation graphique suivante. Nous choisissons donc finalement le modèle $\boldsymbol{\delta}$. Confirmer ce choix à l'aide de la fonction `step`. L'évolution du SIDA du chat est-il lié au sexe du chat ?

. . . . . .