Feuille de Travaux Dirigés n° 3
Analyse de la covariance

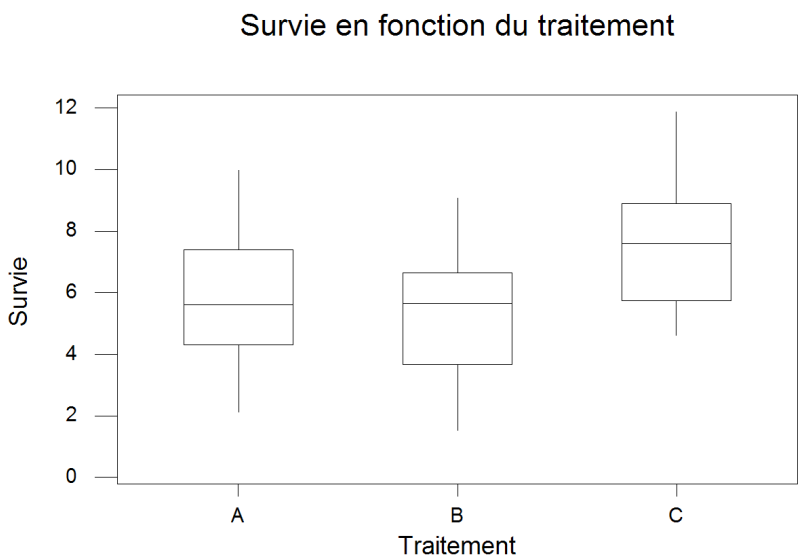
Les exercices 1 et 2 sont tirés du livre *Modèle linéaire : Comparaison de groupes et régression* de B. Prum aux Éditions de l'INSERM

Exercice III.1. Cancer du sein

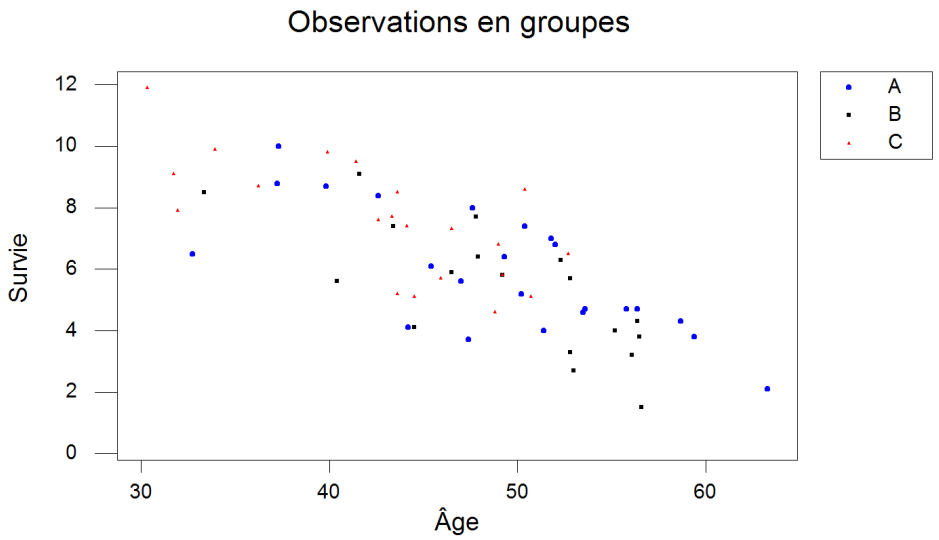
On étudie la durée de survie Y de femmes atteintes de cancer du sein soumises à trois traitements, A, B et C. Ces durées figurent dans le tableau suivant dans les colonnes « Survie » ; on a aussi indiqué l'âge X d'apparition d'un cancer.

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, 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.



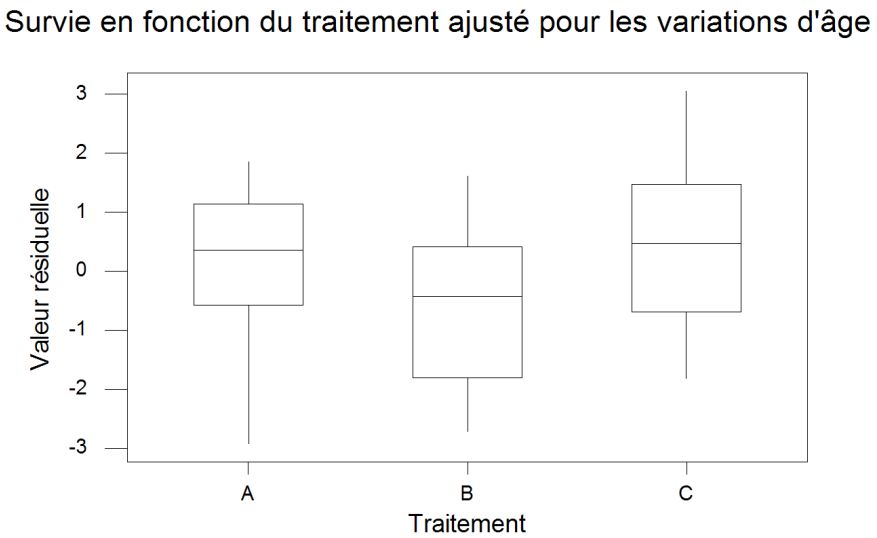
2. Représenter graphiquement les données à l'aide d'un diagramme de la durée de survie en fonction de l'âge où l'on spécifiera de surcroît le traitement reçu.



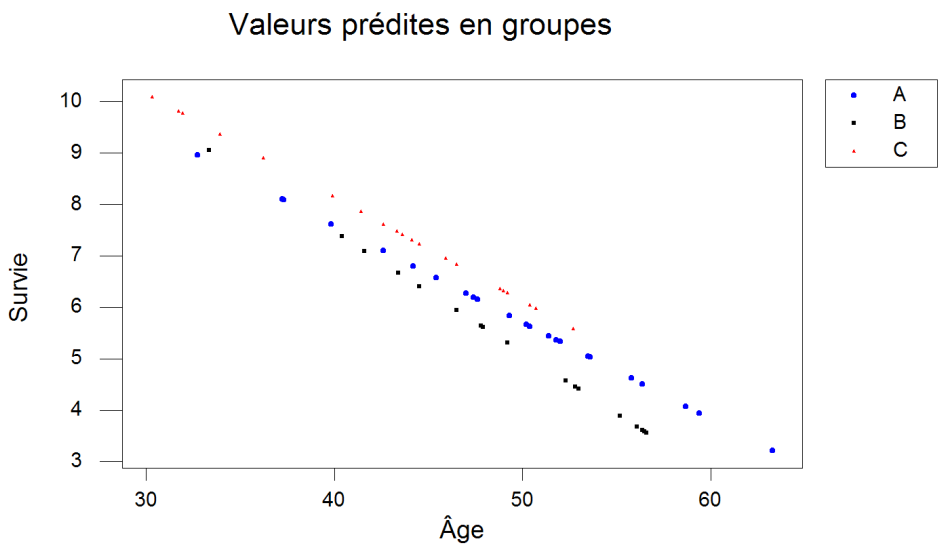
3. On soupçonne un lien entre l'âge d'apparition et la durée de survie, quels modèles peut-on envisager ? Étudier en particulier les modèles qui comportent les termes suivants :

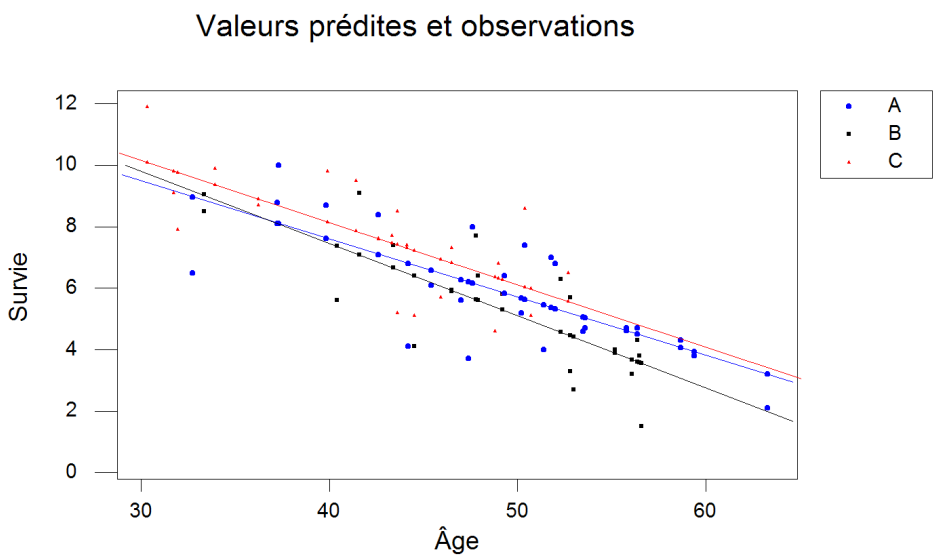
- **a.** Effet linéaire de l'âge.
- **b.** Effet linéaire de l'âge et effets principaux des traitements.
- **c.** Effets linéaires de l'âge différents en fonction des traitements et effets principaux des traitements. Ce modèle permet de savoir si au sein de chacun des groupes le facteur Âge est significatif. On utilisera pour ce modèle un modèle emboîté qui se note Âge(Traitement) et qui signifie que l'on cherche une pente différente pour chaque niveau du traitement.
- **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.

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.** ?



4. Il y a-t-il 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.** et des observations puis à l'aide des valeurs prédites par le modèle du **3.b.** et des observations. Interpréter ces graphiques à l'aide des résultats du test précédent.


Valeurs prédites en groupes
Survie
Âge
A
B
C


Valeurs prédites et observations
Survie
Âge
A
B
C

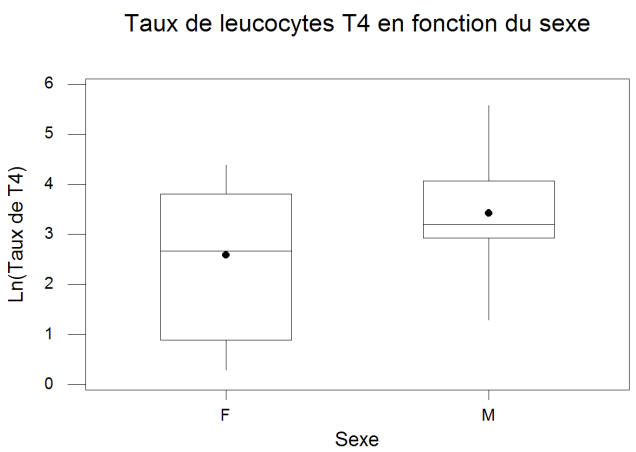
. . . . . .

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



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.
- **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.
- **c.** On considère deux droites de régression, l'une pour les chats mâles, l'autre pour les chats femelles.
- **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.
- **e.** On considère deux droites dont les pentes dépendent 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.
- **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 dépend du sexe du chat comme un écart à une valeur moyenne commune aux deux sexes.

Associer à chacune des formes mathématiques des modèles ci-dessous son interprétation en termes de pentes et d'ordonnées à l'origine qui figure dans la liste ci-dessus.

- $\boldsymbol{\alpha}$. $Y_{i,j} = \beta_0 + \alpha_i + \beta_1 X_{2,i,j} + \epsilon_{i,j}$.
- $\boldsymbol{\beta}$. $Y_{i,j} = \alpha_i + \beta_1 X_{2,i,j} + \epsilon_{i,j}$
- $\boldsymbol{\gamma}$. $Y_{i,j} = \beta_0 + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$
- $\boldsymbol{\delta}$. $Y_{i,j} = \alpha_i + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$
- $\boldsymbol{\epsilon}$. $Y_{i,j} = \beta_0 + \alpha_i + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$
- $\boldsymbol{\phi}$. $Y_{i,j} = \beta_0 + \alpha_i + \beta_1 X_{2,i,j} + \beta_{2,i} X_{2,i,j} + \epsilon_{i,j}$

Chacune de ces équations est valable pour $1 \leq i \leq 2,\ 1 \leq j \leq n_i$ avec $n_1 = 17$ et $n_2 = 15$.

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.

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. On pourra utiliser les indications suivantes :

On commencera par coder séparément la variable « Sexe » à l'aide de deux variables indicatrices l'une pour les chats mâles et l'autre pour les chats femelles. [1].

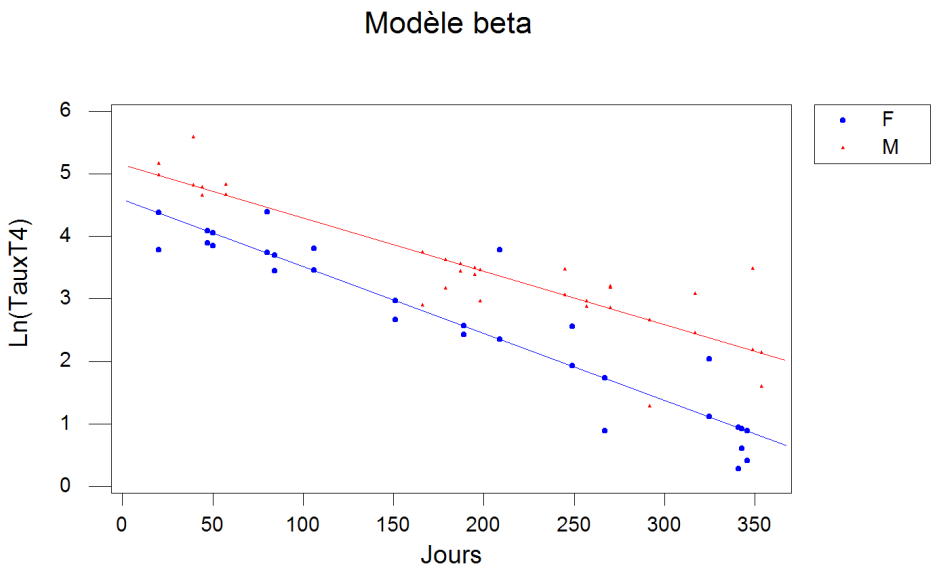
Puis on codera la variable « Jours » à l'aide de deux variables, l'une associée aux jours pour les chats mâles et l'autre aux jours pour les chats femelles[2].

- $\boldsymbol{\alpha}$**.** Modèle linéaire généralisé avec les termes « Sexe » et « Jours », le facteur « Jours » étant modélisé comme covariable.
- $\boldsymbol{\beta}$**.** Faire une régression par rapport aux variables « Mâles », « Femelles » et « Jours » sans ajuster l'ordonnée à l'origine.
- $\boldsymbol{\gamma}$**.** Faire une régression par rapport aux variables « JoursF » et « JoursM » en ajustant l'ordonnée à l'origine.
- $\boldsymbol{\delta}$**.** Faire une régression par rapport aux variables « Mâles », « Femelles », « JoursM » et « JoursF » sans ajuster l'ordonnée à l'origine.
- $\boldsymbol{\epsilon}$**.** Modèle linéaire généralisé avec les termes « Sexe » et « Jours(Sexe) », le facteur « Jours » étant modélisé comme covariable et emboîté dans le facteur « Sexe'.
- $\boldsymbol{\phi}$**.** Modèle linéaire généralisé avec les termes « Sexe », « Jours » et le terme d'interaction « Sexe*Jours », le facteur « Jours » étant modélisé comme covariable.

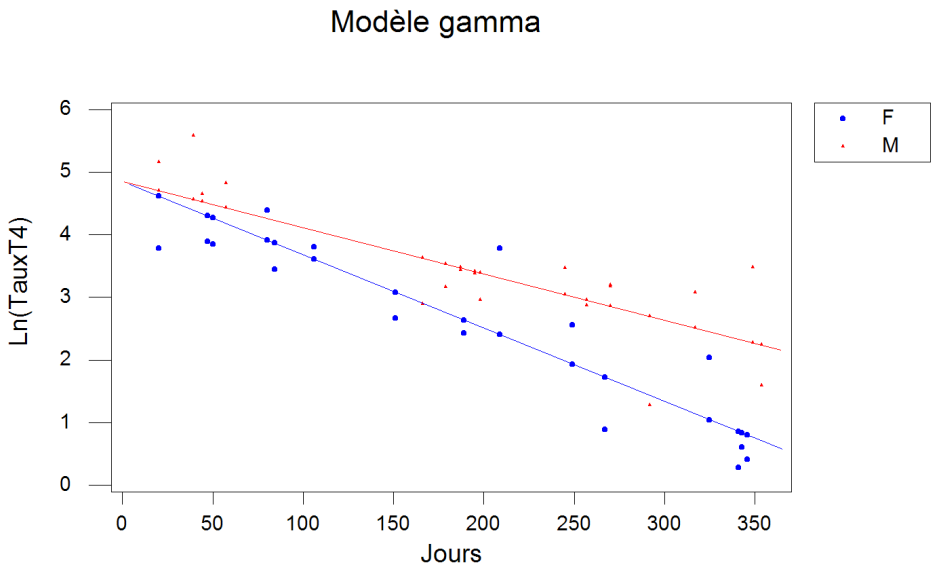
4. On constate que le modèle $\boldsymbol{\beta}$**.** est un des modèles à retenir. L'évolution du SIDA du chat est-il lié au sexe du chat ? Obtenir la représentation graphique associée à ce modèle.

[1] dans la colonne « Male » on notera 1 si le chat est un mâle et 0 si c'est une femelle et dans la colonne « Chatte » on notera 0 si le chat est un mâle et 1 si le chat est une femelle.

[2] La colonne « JoursM » comportera la valeur de la variable « Jours » pour les chats mâles et 0 pour les femelles, la colonne « JoursF » comportera la valeur de la variable « Jours » pour les chats femelles et 0 pour les chats mâles.



5. Le modèle $\boldsymbol{\gamma}$ peut être considéré comme un sous-modèle du modèle $\boldsymbol{\beta}$. En utilisant une somme des carrés de type I, c'est-à-dire séquentielle, 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. L'évolution du SIDA du chat est-il toujours lié au sexe du chat ?



. . . . . .