Analyse du fonctionnement des horloges biologiques

Xiawei She

▶ **To cite this version:**

Xiawei She. Analyse du fonctionnement des horloges biologiques. Méthodologie [stat.ME]. 2014. dumas-01059656

HAL Id: dumas-01059656

https://dumas.ccsd.cnrs.fr/dumas-01059656

Submitted on 1 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse du fonctionnement des horloges biologiques

maître de stage : M. André MALAN

étudiant : Xiawei SHE

Master 1 Statistique

Remerciements

Je souhaite remercier tout particulièrement mon maître de stage, monsieur André Malan, pour l'aide qu'il m'a apportée d'un point de vue mathématique et biologique, pour m'avoir permis d'assister aux manipulations au sein des laboratoires, ainsi que pour l'aide qu'il m'a apportée tout au long de mon stage et plus généralement concernant le fonctionnement de l'horloge biologique ainsi que pour m'avoir permis d'effectuer mon stage au sein de l'INCI.

Je remercie également Jorge Mendoza, chercheur à l'INCI, pour toutes ses explications concernant son sujet de recherche, les protocoles expérimentaux. Je le remercie également pour son accueil chaleureux, sa patience, sa gentillesse et son soutien.

Je remercie toutes les autres travailleurs du laboratoire de neurobiologie des fonctions rythmiques pour leur accueil.

TABLE DES MATIERES

Remerciement ……...2

I. Présentation de l'institut des neurosciences cellulaires et intégratives……..4

II. Présentation de sujet de ce stage……...5

III. Sélection de données……...7

IV. Algorithme appliqué au traitement des donnees……..8

V. Ajustement des données avec les modèles complets……...11

VI. Traitement de colinéarité des paramètres et Simplification du modèle..……..…..14

VII. Comparaison entre le modèle complet et le modèle simplifié……..……...…18

VIII. Recherche des caractéristiques communes au sein de chaque groupe…..…….21

IX. Comparaison les caractéristiques entre les groupes…….…..……...38

X. Conclusion…….…….43

I. Présentation de l'institut des neurosciences cellulaires et intégratives (l'INCI)

J'ai effectué mon stage à l'institut des neurosciences cellulaires et intégratives (INCI) ,qui est un laboratoire commun et du CNRS (UPR 3212) et aussi une partenaire de l'université de Strasbourg. La neuroscience est une domaine concernant le organisme et le fonctionnement du système nerveux. L'INCI s'intéresse aux mécanismes cellulaires et moléculaires impliqués dans la régulation de système nerveux qui est associé à l'amélioration de la qualité de la vie humaine.

L'INCI est un institut de recherche composé de neuf équipes par une approche multidisciplinaire (génomique, protéomique, électrophysiologique, physiologique et comportementale). Et les neufs équipes se concernent sur trois thèmes majeurs de Neurophysiologie (les rythmes biologiques, la neurosécrétion et la nociception).

L'INCI se divise par trois départements :

1) Département de physiologie des réseaux neuronaux. Il y a deux équipes dans le département. Le département se concerne la recherche de la propriétés et plasticité de la neurotransmission, sécrétion neuro-endocrinienne, traitement des informations et organisation fonctionnelle des microcircuits neuronaux.

2) Département de nociception et douleur. Il y a deux trois équipes dans le département. Le département se concerne la recherche du réseaux de neurones et signalisation nociceptive dans la moelle épinière, approches moléculaires du contrôle de la douleur, mécanisme et traitement de la douleur chronique.

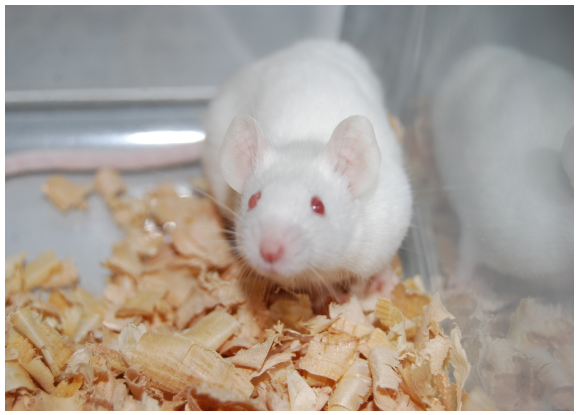
3) Département de neurobiologie des fonctions rythmiques. Il y a quatre équipes dans le département. Le département se concerne la recherche de la genèse et synchronisation des signaux circadiens dans le système nerveux, mélatonine et rythmes saisonniers, cycle circadien et physiopathologie de la rétine, contrôle circadien et homéostatique du sommeil.

Axes thématiques	Équipes
Communication et réseaux de neurones	• **Trafic membranaire dans les cellules du système nerveux** Responsables : Stéphane GASMAN & Nicolas VITALE • **Physiologie des réseaux de neurones** Responsables : Philippe ISOPE & Bernard POULAIN
Nociception et douleur	• **Signalisation nociceptive dans la moelle épinière** Responsable : Rémy SCHLICHTER • **Déterminants moléculaires de la douleur** Responsable : Pierrick POISBEAU • **Douleur chronique : approche anatomo-fonctionnelle et traitement** Responsable : Michel BARROT
Neurobiologie des rythmes	• **Rythme, vie et mort de la rétine** Responsables : David HICKS & Frank PFRIEGER • **Régulation des horloges circadiennes** Responsable : Etienne CHALLET • **Mélatonine et rythmes saisonniers** Responsable : Valérie SIMONNEAUX • **Lumière, rythmes et homéostasie du sommeil** Responsable : Patrice BOURGIN

II Présentation de sujet du stage

Pour réaliser ce stage, j'étais intégré dans le département de neurobiologie des fonctions rythmiques. Et ma étude statistique concerne l'activité rythmiques sur dans le cerveau d'animal.

Les organismes vivants ont des rythmes biologiques qui peut être observé en expression comportementale ou physiologique. Le rythme biologique principale est le cycle de réveil/sommeil exprimant un motif circadien (c'est à dire que le rythme admet un cycle de 24 heures). Les processus physiologiques montrent un motif circadien par un processus d'élevage. Normalement, l'ingestion de nourriture des humaines a lieu pendant la journée, mais l'ingestion de nourriture d'animal rongeur expérimental a lieu pendant la soirée. Pour les deux cas (humaine et rongeur), l'ingestion de nourriture est associée au cycle d'activité normal.



L'ingestion de nourriture est réglée au moins par deux facteurs concernés dans le cerveau : un homéostatique mécanisme et un hédonistique (ou récompense) mécanisme. La surconsommation de nourriture savoureuse qui est modulée par un hédonistique mécanisme, est un facteur qui est contribue à l'obésité. Les animaux sous condition du choix libre au nourriture savoureuse riche en lipides et sucre deviennent rapidement hyperphagique. C'est qui nous intéresse, c'est que l'ingestion de nourriture savoureuse a lieu principalement au moment de la période d'endormi qui suggère que le hédonistique mécanisme dans le cerveau est plus sensible pendant la période.

L'objective de ce projet est de déterminer si la relation fonctionnelle entre l'horloge circadienne et le mécanisme hédonistique central est impliqué grâce au augmentation d'ingestion de nourriture savoureuse pendant la période d'endormi, ainsi de comprendre le mécanisme neural moléculaire.

Pour analyser l'effet de différente sorte de régime, on a crée deux groupes :

Groupe 1 : Groupe contrôlée
Les animaux sous condition de l'eau et nourriture ordinaire.

Groupe 2 : Groupe fcHFHS
Les animaux sous condition du choix libre au nourriture savoureuse riche en lipides et sucre. Les animaux de ce groupe ont quatre choix comme nourriture :l'eau ordinaire, l'eau sucré de 10% (concentration le plus acceptable pour les souris), des pastilles de nourriture ordinaire et des pastilles de nourriture grasse.

L'expérimentation a duré 6 semaines. Le poids du corps des souris et l'ingestion de nourriture sont mesurés tous les semaines.

Pendant la procédure de expérimentation, on a découvert que les souris ont toujours la préférence de l'ingestion de la nourriture grasse et l'eau sucré. En outre, les calories l'ingestion de la nourriture grasse et l'eau sucré sont absorbées principalement pendant la période de soirée où les souris devaient être endormi.

Afin d'analyser l'expression de la protéine PER2 dans le cerveau, on utilise des souris mutants qui ont le luciférase accouplé au gène de l'horloge PER-2. Ce gène possède une expression circadienne montrant une crête au début de la période de soirée. On peut observer les différentes parties du cerveau de souris en utilisant le système de photomultiplicateur (PMT) qui est capable de détecter les photons libéré par l'oxydation du luciférase. L'équation est comme suivante

$$luciférase + ATP + O_2 \xrightarrow{Mg^{2+}} Oxyluciférase + AMP + PPi + CO_2$$

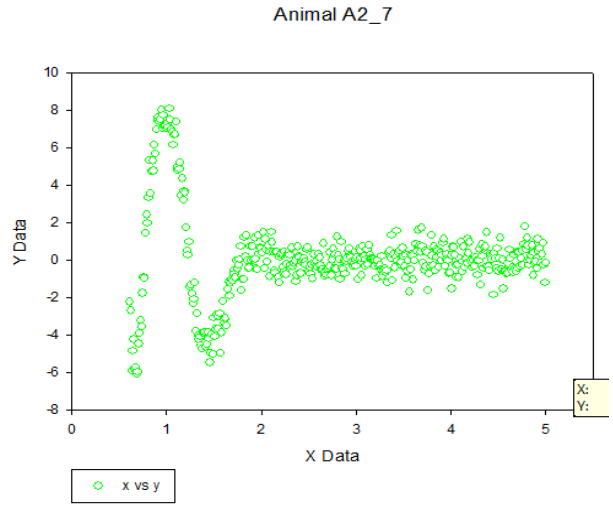
La lumière lancée par le tissu est acceptée par le photomultiplicateur et les données sont enregistrées par l'ordinateur. Donc on peut analyser les données obtenues et étudier les modèles d'oscillation du tissu qui nous intéresse.

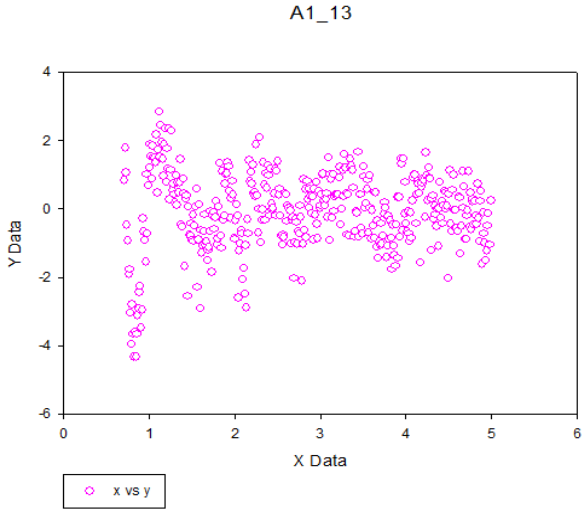
Parce que les conditions d'expérimentation dans le laboratoire(température, lumière ou fermeture e la porte etc) changeait extrêmement faible, on a obtenu des données tordues qui devait être enlevées avant l'analyser.

III. Sélection des échantillons

Au début, il y a eu 14 au total, 7 animaux par groupe. Avant l'analyse, j'ai tracé les nuages des données de tous les échantillons. Je suis prévenu que le fonctionnement des horloges dans le cerveau de l'animal (souris) est périodique et amortie. Je suis obligé de déterminer la faisabilité d'analyse de données en observant les nuages des données.

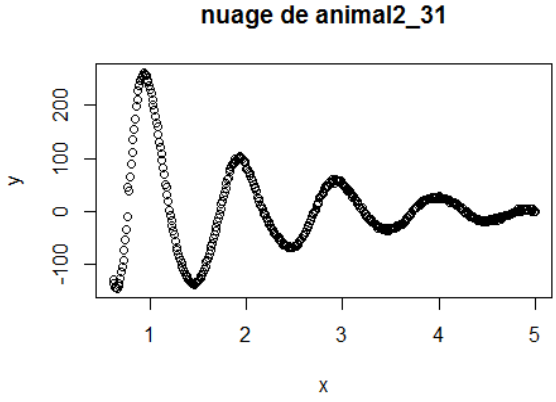
Par exemple, on a le nuage d'échantillon A2_7. On peut voir que l'oscillation ne dure pas assez long temps (environ 1 unité du temps, converti en format d'heure, c'est 24.04 heures). S'il ne dure pas assez long temps, le résultat qu'on aura obtenu ne serrent pas fiable statistiquement .Et on peut aussi trouver un échantillon suivant tel que l'on ne peut pas voir l'oscillation intuitive.

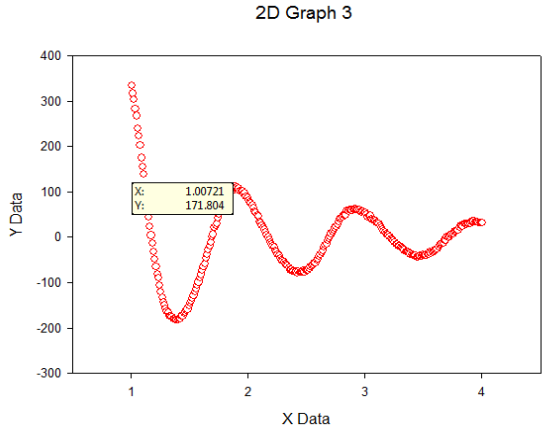




Il n'y a vraiment pas de sens d'étudier les situations comme précédentes, donc j'ai enlevé tous les échantillons <tordus> comme les échantillons précédentes. J'ai sélectionné les échantillons tels que l'on peut voir oscillation amortie assez apparente et il dure assez long temps.

Un autre problème est qu'il faut un peu de temps aux cellules en culture pour s'adapter aux conditions de culture. En outre, on ne peut pas voir oscillation assez apparente ou l'oscillation sera hyper faible après le temps x=4 pour quasiment toutes les courbes tracées. Donc je suis conseillé d'enlever les points enregistrés tels que le temps x est inférieur à 1 ou supérieur à 4. Ils ne seront pas pris en compte à l'analyse. Les courbes sont donc analysées à partir du temps x=1 au x=4 pour tous les échantillons.

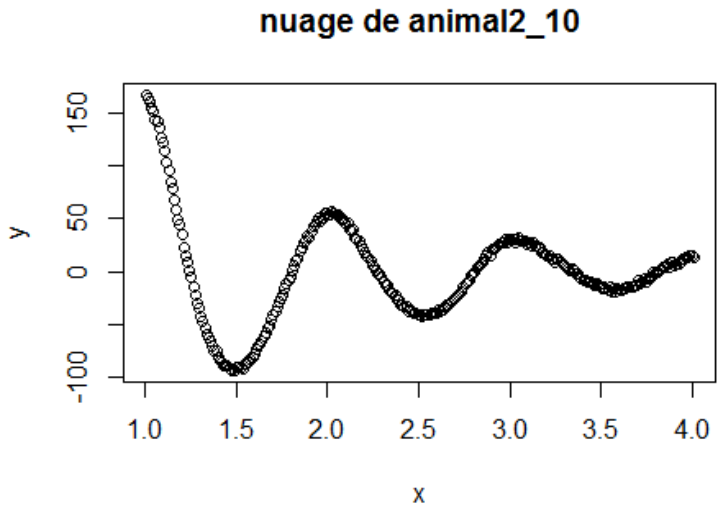




IV. Algorithme appliqué au traitement des données

1. Régression non-linéaire

Tout d'abord, on trace le nuage des données (par exemple échantillon animal25) où l'abscisse est le temps et l'ordonnée représente les données de lumière qui vient de la luciférase. Le graphe est suivant,



Intuitivement, on ne peut pas trouver un modèle linéaire pour l'ajustement. Donc je suis exigé de chercher un modèle non-linéaire s'écrivant

$$y_i = f(t_i, \Theta) + \varepsilon_i$$

où i={0,1…n},
- ε est terme des erreurs
- Θ est l'ensemble des paramètres
- y sont des observations
- f est une fonction non-linéaire

Le modèle non-linéaire classique admet une ε qui sont de la loi normal centré et ont une variance commune σ². Et les erreurs sont non-corrélés, donc on a toujours

$$E(y_i) = f(t_i, \Theta)$$

Pratiquement, pour trouver une fonction $f(t_i, \Theta)$ le plus convenable (le modèle se produit des résidus minimaux), on peut toujours chercher somme des carrés des résidus $\sum_{t=0}^{n} (y_i - f(t_i, \Theta))^2$

2. Algorithme de Gauss-Newton

Pour chercher le modèle statistique non-linéaire, on a utilisé algorithme de Gauss-Newton qui est une méthode universelle pour résoudre des problèmes non-linéaires. Cet algorithme nous donne une résolution de moindres carrés. C'est à dire que l'on a obtenu un résultat de la minimisation d'une somme des fonctions.

$$S(\Theta)=f(x)=\sum_{t=0}^{n}(y_i-f(t_i,\Theta))^2$$

En interprétant cet algorithme , je suppose qu'il y a n fonctions $h_i=y_i-f(t_i,\vartheta)$,et θ est le vecteur des paramètres, $\vartheta=(\vartheta_1,\vartheta_2,\ldots,\vartheta_m)$. On a n≥m. Le but de cet algorithme est de trouver la valeur minimale de $S(\vartheta)=\sum_{i=0}^{n}(h_i(\vartheta))^2$.

L'algorithme est basé sur un développement en série de Taylor au voisinage des valeurs initiales des paramètres. Par exemple, dans le projet, on a travaillé sur un modèle avec 9 paramètres $\Theta=(e,g,h,i,a,d,b,c)^T$ .Cet algorithme procède des itérations avec un vecteur initial de paramètre notera $\Theta_0=(e_0,g_0,h_0,i_0,a_0,d_0,b_0,c_0)$. Le modèle s'écrit :

$$f(x,\Theta)=f(x,\Theta_0)+\frac{\partial f(x,\Theta_0)}{\partial e}(e-e_0)+\frac{\partial f(x,\Theta_0)}{\partial g}(g-g_0)+.....+\frac{\partial f(x,\Theta_0)}{\partial c}(c-c_0)$$

On peut exprimer l'itération comme suivant :

FAIRE k = 0,1,2..... :

Calculer le vecteur $f_k(x,\Theta_k)$ et la matrice $J_k(x,\Theta_k)$

Calculer le vecteur de paramètres $\Delta\Theta=([J_k(x,\Theta_k)]^T[J_k(x,\Theta_k)])^{-1}[J_k(x,\Theta_k)]^T f_k(x,\Theta_k)$

Calculer $SSR=\|Y-f(x_k,\Theta_k)\|_2^2$

Calculer $\Theta_{k+1}=\Theta_k+\Delta\Theta$ et k = k+1

ROCOMMENCER LE BOUCLE JUSQUE À SSR converge vers une valeur

END

Ici, $J_k(x,\Theta_k)$ est la matrice jacobienne de la fonction $h(\Theta_k)=\begin{pmatrix}h_0(\Theta_k)\\ \vdots \\ h_n(\Theta_k)\end{pmatrix}$ par rapport à Θ_k .

Il existe comme même des autres algorithmes pour le régression non linéaire. Pour ce projet, on a choisi algorithme de Gauss-Newton, qui correspond à la fonction <nls()> sous R.

V. Ajustement des données avec les modèles complets

Après la sélection des données, on a deux groupes de données à analyser, et chaque groupe contient 4 échantillons. Tous d'abord, je vais présenter tous les modèles de chaque échantillon et la qualité d'ajustement. Deuxième, je proposerai d'une méthode de simplifier les modèles et les modèles simplifiés.

On a tracé les nuages des points des données des deux groupes. Et on trouve que toutes les courbes sont de la forme d'une fonction sinusoïdale amortie approximativement.

Intuitivement, la courbe n'est pas de la forme sinusoïdale amortie <standard>, et on ne peut pas y ajuster avec une seule fonction sinusoïdale amortie $y_i = f(x_i) = a * \exp(-x_i/d) * \sin(\frac{2*\pi*x_i}{b} + c)$, parce qu'on n'est pas sûr que si l'oscillation forme une ligne de base non horizontale, et la régression se produit des grandes erreurs.

Plus précisément, on peut chercher un modèle mixte avec une fonction polynomiale admettant plusieurs paramètres et une fonction sinusoïdale amortie. On prend le premier groupe comme un exemple. Pour ajuster les données, on a travaillé sur le modèle comme suivant :

$$y_i = f(x_i) = y0 + e * x_i + g * x_i^2 + h * x_i^3 + a * \exp(-x_i/d) * \sin(\frac{2*\pi*x_i}{b} + c)$$

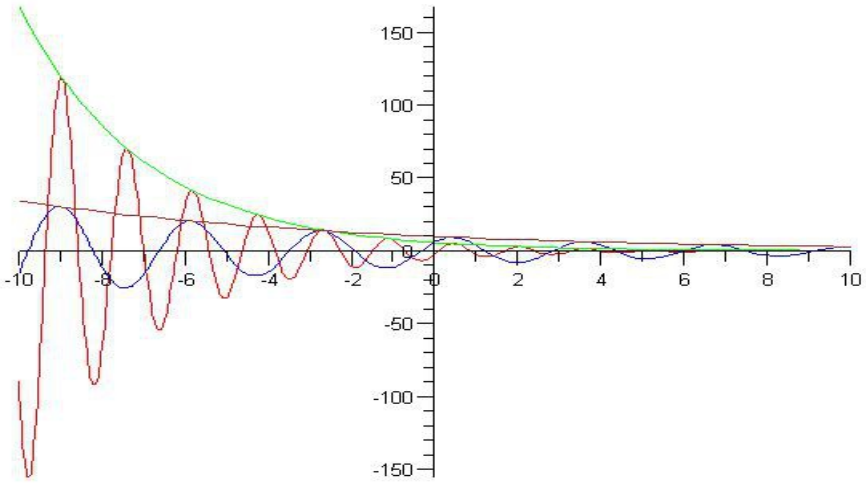
où
y représente les données associée à l'activité de cerveaux d'animal
x représente le temps en format numérique. Par exemple, quelque soit i, la différence dans le modèle correspond bien à 15 minutes en réalité.

Le paramètre <b> représente la période de l'oscillation. rappelons que la période T est égal à $T = 2 \cdot \pi / \omega = 2 \cdot \pi / (2 \cdot \pi / b) = b$ où $2 \cdot \pi / b$ b est la pulsation et <c> représente la phase. Et les paramètres de la partie polynomiale (y0, e, g, h) n'ont pas de signification biologique. Les paramètres<b>,<c> et <d> sont celles dont les estimations nous intéressent le plus, car les études biologiques portes sur la période, la phase et le niveau d'amortissement de l'oscillation dans les cerveaux.

Ici, b et c sont au format numérique, mais l'unité de la période et la phase doit être convertie en format du temps réel (en heure ici). Donc, dès que l'on a obtenu la valeur de b ou c, on la converti au format du temps en unité heure par l'équation suivante :

$$Période = \frac{b \times 15}{0.0104 \times 60} \quad \text{Ou} \quad Phase = \frac{c \times 15}{0.0104 \times 60}$$

Le paramètre d décrit l'effet d'amortissement de l'oscillation. La valeur est plus petite. L'amplitude d'oscillation diminue plus rapidement, c'est l'envers pour la valeur plus grand. Par exemple dans le graphe suivant, la courbe rouge admet un d=6 et la courbe bleu admet un d=8.



Le paramètre a décrit l'amplitude de l'oscillation.

Par exemple, pour le premier groupe, on a ajusté les données avec le modèle

$$y_i = f(x_i) = y0 + e * x_i + g * x_i^2 + h * x_i^3 + a * \exp(-x_i / d) * \sin\left(\frac{2 * \Pi * x_i}{b} + c\right)$$

Pour l'échantillon A1-4, les estimations des paramètres :

Paramètres	Estimations	Erreurs Standards	t-value	p-value
a	188.7293	3.6705	50.1875	<0.0001
b	1.0443	0.0022	480.3024	<0.0001
c	2.2738	0.0242	93.8783	<0.0001
d	1.3689	0.0206	66.4895	<0.0001
y0	94.5436	6.3419	14.9078	<0.0001
e	-131.8277	8.2162	-16.0448	<0.0001
g	54.6413	3.3332	16.3928	<0.0001
h	-6.9154	0.4279	-16.1605	<0.0001

Le coefficient de détermination R-carré =0.9901 qui est assez grand. Donc le modèle bien ajuste les données.

Le résultat de test de normalité des résidus :

Shapiro-Wilk normality test

data: residuals(model1_4)
W = 0.9912, p-value = 0.08293

Le p-value est supérieur à 0.05, dont le test n'est pas significatif, on garde l'hypothèse nulle. Donc l'hypothèse de normalité de la régression non-linéaire est vérifiée.

Le résultat de test de homoscédasticité des résidus :

Bartlett test of homogeneity of variances

data: list(residuals(model1_4), t1_4$x)
Bartlett's K-squared = 361.8629, df = 1, p-value < 2.2e-16

Le p-valeur est beaucoup inférieur à 0.05, le test est significatif. On rejette l'hypothèse nulle. L'hypothèse de homoscédasticité des résidus n'est pas vérifiée.

On peut savoir la période de l'oscillation T = b = 1.0443. On a converti le temps au format numérique. Toutes les 15 minutes correspondent à 0.0104. Donc T = 1506.20 mins = 25.1 heures=1.06 jour. La phase de l'oscillation est c = 2.2738 = 3279.51 mins = 54.66 heures= 2.17 T

Les modèles de régression non-linéaire et des résultats d'analyse sont en Annexe 1.

Donc on peut conclure que tous les modèles ont bons résultat d'ajustement, car les coefficients de détermination des modèles sont tous assez grand et supérieur à 0.9. Les tests de normalité des résidus sont presque tous non significatifs sauf l'animal A2_13. C'est à dire que la normalité des résidus des autres régressions est vérifiée.

Mais au niveau de l'homoscédasticité, presque tous les tests sont significatifs. J'ai essayé de résoudre ce problème en appliquant des transformations sur la variable .On lui a appliqué plusieurs fonctions (logarithme, racine, puissance), mais on n'a pas réussi à obtenir l'homogénéité des résidus. On ne peut quasiment pas trouver un modèle de régression qui vérifie l'homoscédasticité des résidus.

VI. Traitement de colinéarité des paramètres et Simplification du modèle

On a tenu compte du fait qu'on a d'excellents indicateurs pour l'ensemble des régressions. On a en effet des valeurs très élevées de R^2 et de la statistique de Fisher F. On n'a que très rarement des valeurs de coefficient de détermination inférieures à 0.90 et les valeurs de F sont presque toujours supérieurs à 100, parfois même supérieures à 1000. On constate graphiquement que les ajustements sont très bons et on considère donc que les régressions peuvent être validées.

Cependant le diagnostique des régressions révèle un autre problème : celui des valeurs trop élevées des facteurs d'inflation de la variance (VIF). En statistique, la valeur de VIF est quantité mesurant le niveau de multi-colinéarité d'analyse de régression par la méthode des moindres carrés ordinaires (MCO). Il indique le niveau d'inflation de la variance de paramètre grâce à manque de l'indépendance. Donc une valeur de VIF trop élevé révèle une régression « instable » et douteux.

On peut prendre l'échantillon de l'animal A2_10 comme un exemple. On peut voir que les valeurs de VIF des paramètres de la partie polynomiale (y0, e, g, h, i) sont tous supérieur à 10^4, mais les valeurs de VIF des paramètres de la parie sinusoïdale amortie (a, b, c, d) sont tous inférieur à 50. On peut dire que le si l'existence de colinéarité dans la partie polynomiale est beaucoup plus significative, car les valeurs de VIF des paramètres de la partie polynomiale sont beaucoup plus important que celles de la parie sinusoïdale amortie.

Paramètres	VIF
a	17.5814
b	19.2414
c	25.7617
d	14.3048

Paramètres	VIF
y0	17385.3627
e	378033.9960
g	1455793.5272
h	1191409.5703
i	138261.0309

Donc la chose que l'on doit faire dans un premier temps, c'est d'essayer d'éliminer un ou des paramètres de la partie polynomiale pour abaisser la colinéarité entre les paramètres et simplifier le modèle.

Avant d'éliminer les paramètres du modèle, on a besoin d'identifier les corrélations entre les paramètres en trouvant la matrice de corrélation. Toujours dans le cas du même échantillon, on obtient la matrice de corrélation suivante :

	a	b	c	d	y0	e	g	h
b	-0.297							
c	-0.348	0.966						
d	-0.959	0.277	0.323					
y0	-0.658	0.550	0.624	0.563				
e	0.643	-0.526	-0.593	-0.546	-0.997			
g	-0.625	0.508	0.569	0.527	0.988	-0.997		
h	0.606	-0.497	-0.551	-0.508	-0.976	0.990	-0.998	
i	-0.588	0.489	0.537	0.491	0.962	-0.980	0.992	-0.998

Rappelons l'équation sur laquelle on travaille :

$$y = f(x) = y0 + e*x + g*x^2 + h*x^3 + i*x^4 + a*\exp(-x/d)*\sin\left(\frac{2*\pi*x}{b} + c\right)$$

D'après le tableau précédent, on peut voir que dans l'équation, les paramètres y0, e, g, h, i sont fortement corrélés entre eux. C'est également le cas des paramètres a et d et des paramètres b et c.

On a identifié les paramètres responsables de la colinéarité. La multilinéarité se traduit la fois au niveau de la matrice de corrélation entre paramètres et au niveau du VIF.

On peut désormais simplifier l'équation en éliminant les paramètres superflus afin de réduire le phénomène de multi-colinéarité. Cela se traduira par une diminution de VIF.

Pour l'élimination des paramètres du modèle, j'ai appliqué à tous les données la méthode descendante. Les étapes sont suivantes :

1 : Calculer la régression pour le modèle incluant toutes les k paramètres.

2 : Effectuer un test de Student pour chacune des paramètres. Éliminer la paramètre le moins significative du modèle.

3 : Recommencer le processus avec un paramètre en moins jusqu'à le modèle est optimale (les valeur de VIF ne sont plus important et égaux à dizaine)

Ici, on a les valeurs de statistique de student t pour tous les paramètres de modèle pour l'échantillon A2_10 :

Paramètres	Estimations	t-value
a	287.7983	63.2056
b	1.0519	633.8633
c	2.0192	107.7813
d	1.2955	93.0107
y0	482.9267	22.6275

e	-812.9642	-21.6686
g	479.8697	20.4017
h	-120.0508	-19.1364
i	10.8724	17.9987

On trouve que la valeur de t² de paramètre i est la plus petite, donc le test pour i est le moins significative du modèle. Donc je l'élimine. Et plus je recommence une régression sans le paramètre i. J'ai noté les informations (estimations, la valeur de t et VIF) dans le tableau suivant :

Paramètres	Estimations	t-value	VIF
a	340.8641	53.0453	12.8621
b	1.0399	486.3150	15.6285
c	1.8665	80.3416	19.4689
d	1.1817	77.2523	11.8544
y0	113.7713	12.8481	1360.5915
e	-151.220	-13.2779	15836.5952
g	59.7625	13.0140	25227.1037
h	-7.2880	-12.4272	4733.4599

Cette fois, on a obtenu un résultat meilleur, car les valeurs de VIF est abaissé. Le coefficient de détermination R-carré = 0.9929. Mais les valeurs de VIF sont encore beaucoup supérieures à 100. Donc on sélectionne un autre paramètre à éliminer. On trouve que la valeur de t² de paramètre h est la plus petite, donc le test pour h est le moins significative du modèle. Donc je l'élimine. Et plus je recommence un régression sans le paramètre h.

Paramètres	Estimations	t-value	VIF
a	382.9943	47.9756	10.9030
b	1.0368	386.2627	16.0899
c	1.7985	64.1359	19.0330
d	1.1032	70.7406	10.6071
y0	9.7312	2.6854	148.5130
e	-12.7676	-4.2996	701.7144
g	2.9667	5.2421	249.7308

Cette fois, même si on élimine un autre paramètre h, le coefficient de détermination R-carré = 0.9890. Mais les valeurs de VIF sont encore grand, et un problème se produit, c'est que le test de student pour le paramètre h n'est plus significatif. Donc on refait la procédure sans le paramètre h.

Finalement, on peut trouver que le modèle de régression non linéaire est capable d'ajuster les données avec des valeurs de VIF plus petite, mais garder la valeur du coefficient de determination R-carré et la statistique de fisher F plus grande.
On obtenu les résultats suivants :

Paramètres	Estimations	t-value	VIF
a	391.1580	42.7735	10.2005
b	1.0364	397.9689	11.1876
c	1.7997	71.3039	11.4808
d	1.0893	63.6290	9.9146

Donc on peut ajuster les données par le modèle suivant :

$$y = f(x) = a * \exp(-x/d) * \sin(\frac{2 * \pi * x}{b} + c)$$

Le coefficient de détermination R-carré = 0.9850 qui est assez grand. Les valeurs de VIF sont environs 10 pour tous les paramètres. Donc on peut dire qu'il n'existe pas de multi-colinéarité grave entre les paramètres. On peut aussi conclure que l'oscillation forme donc une ligne de base horizontale. Il n'y a plus colinéarité entre b et d, parce que le coefficient de corrélation est égale à -0.145.

Ce modèle est plus fiable.

Les résultats de simplification pour les autres échantillons sont en Annexe 2.

VII.Comparaison entre le modèle complet et le modèle simplifié

Après la suppression des paramètres superflus, on a établi une nouvelle équation pour chaque échantillon. On a donc de nouvelles estimations des paramètres. Dans cette partie, on concerne au changement de la qualité de ajustement, la période (correspondant au paramètre <b>), la phase (correspondant au paramètre <c>) et le coefficient d'amortissement (correspondant au paramètre <d>) entre le modèle complet et modèle simplifié.

Prenons l'exemple de l'échantillon A2_1. On a obtenu une nouvelle équation par la méthode de la régression descendante. Avant, on obtenait une estimation pour les 9 paramètres pour l'équation

Paramètres	Estimations	Erreurs Standards	t-value	p-value	VIF
a	339.9023	6.1587	55.1908	<0.0001	10.6400
b	0.9820	0.0016	624.6827	<0.0001	13.3094
c	2.2340	0.0215	104.0561	<0.0001	17.9806
d	1.6716	0.0261	64.1492	<0.0001	10.6359
y0	612.1835	37.1759	16.4672	<0.0001	10409.1828
e	-1105.1921	66.5739	-16.6010	<0.0001	233957.8973
g	699.6393	42.4090	16.4974	<0.0001	927470.5211
h	-185.9650	11.4657	-16.2193	<0.0001	777747.6511
i	17.6974	1.1168	15.8460	<0.0001	92107.1974

On peut savoir la période de l'oscillation T = b = 0.9820. On a converti le temps au format numérique. Toutes les 15 minutes correspondent à 0.0104. Donc T = 1416.35 mins = 23.6 heures. La phase de l'oscillation est c = 2.2340 = 3222.12 mins = 53.70 heures Le coefficient d'amortissement d est égal à 1.6716. Les résidus d'ajustement vérifient la normalité. Le coefficient de détermination est assez grand (0.9918). Mais les VIF sont excessifs.

Après suppression des paramètres superflus, on obtient l'équation suivante :

$$y = f(x) = g * x^2 + a * \exp(-x/d) * \sin\left(\frac{2 * \pi * x}{b} + c\right)$$

.

Les estimations des paramètres :

Paramètres	Estimations	Erreurs Standards	t-value	p-value	VIF
a	362.6203	8.3856	43.2434	<0.0001	9.0641
b	0.9730	0.0017	586.3396	<0.0001	7.8203
c	2.0670	0.0197	105.1558	<0.0001	7.8436
d	1.5989	0.0310	51.5680	<0.0001	9.0902
g	0.2246	0.0631	3.5610	0.0004	1.0329

On peut savoir la période de l'oscillation T = b = 0.9730. On a converti le temps au format

numérique. Toutes les 15 minutes correspondent à 0.0104. Donc T = 1403.37 mins = 23.39 heures. La phase de l'oscillation est c = 2.0670 = 2981.25 mins = 48.69 heures. Le coefficient d'amortissement d est égal à 1.5989. Bien que les résidus ne sont plus vérifient la normalité, le coefficient de détermination est assez grand (R-carré =0.9835), et les valeurs de VIF sont diminuées (elles sont toutes inférieur à 10).

En effet, les modèles simplifiés de tous les échantillons ne sont pas absolument identiques. On peut voir que presque toutes les termes des résidus ne sont plus vérifiées la normalité et l'homoscédasticité. Mais on atteint notre objective. Après la simplification du modèle, on a gardé les valeurs de coefficient de détermination R-carré et de la statistique de fisher le plus grand (toutes les valeurs de coefficient de détermination R-carré sont supérieures à 0.9), en même temps, on a bien diminué les valeurs de VIF (tous les valeurs de VIF sont environ 10 ou encore plus petite). Le niveau de multi-colinéarité du modèle est abaissé remarquablement. On constate également que la moyenne de l'erreur type est diminuée pour tous les groupes. Les modèles sont plus fiables.

On peut comparer les périodes (correspondant au paramètre <b>), les phases (correspondant au paramètre <c>) et les coefficients d'amortissement (correspondant au paramètre <d>) entre le modèle complet et modèle simplifié.

Avant la suppression des paramètres superflus, on obtient les résultats suivants pour chaque groupe :

Groupe 1: Les souris sont nourri avec des aliments gras

échantillon	Période (en heures)	Phase (en heures)	Coefficient d'amortissement
A1_4	25.10	54.66	1.3689
A1_16	25.24	57.42	1.4356
A1_19	25.03	65.75	1.6188
A1_22	24.54	56.31	1.6375

Groupe 2: Les souris sont nourri avec des aliments normaux

échantillon	Période (en heures)	Phase (en heures)	Coefficient d'amortissement
A2_1	23.60	53.70	1.6716
A2_10	25.29	48.54	1.2955
A2_13	25.17	62.47	1.1390
A2_31	24.26	47.88	1.5047

Après la suppression des paramètres superflus, on obtient les résultats suivants pour chaque groupe :

Groupe 1: Les souris sont nourri avec des aliments gras

échantillon	Période (en heures)	Phase (en heures)	Coefficient d'amortissement
A1_4	24.97	52.45	1.2600

A1_16	25.43	58.67	1.3591
A1_19	25.23	67.58	1.4252
A1_22	25.08	62.23	1.3852

Groupe 2: Les souris sont nourri avec des aliments normaux

échantillon	Période (en heures)	Phase (en heures)	Coefficient d'amortissement
A2_1	23.39	49.69	1.5989
A2_10	24.91	43.26	1.0893
A2_13	24.10	58.65	1.0088
A2_31	23.92	42.54	1.3516

Donc, on obtient des périodes moyennes assez proches de celles calculées avant la suppression des paramètres.

Pour la phase, on ne peut pas voir des différences remarquables entre les modèles complets et les modèles simplifiés, sauf l'échantillon A2_13 du groupe 2, on avait une période égal à 62.47 heures. On a désormais une période égal à 48.02, qui est une valeur bien plus proche de celles des autres échantillons du même groupe.

Pour les coefficients d'amortissement, on peut voir les valeurs des modèles complets sont un peu plus grand que celles des modèles simplifiés, mais les différences ne sont pas vraiment décidées.

VIII. Recherche des caractéristiques communes au sein de chaque groupe

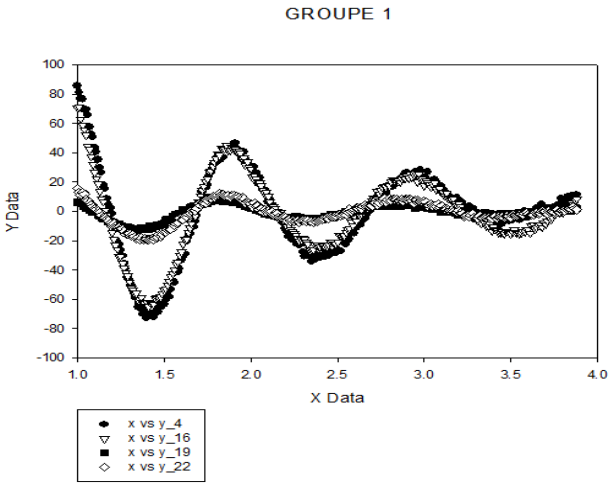
L'objectif de cette partie est de mettre en évidence l'existence des caractéristiques entre les différents échantillons d'un même groupe, c'est-à-dire que les quantités observées (les périodes (correspondant au paramètre <b>), les phases (correspondant au paramètre <c>) et les coefficients d'amortissement (correspondant au paramètre <d>)) sont globalement assez proche pour qu'on considère que le groupe est cohérent.

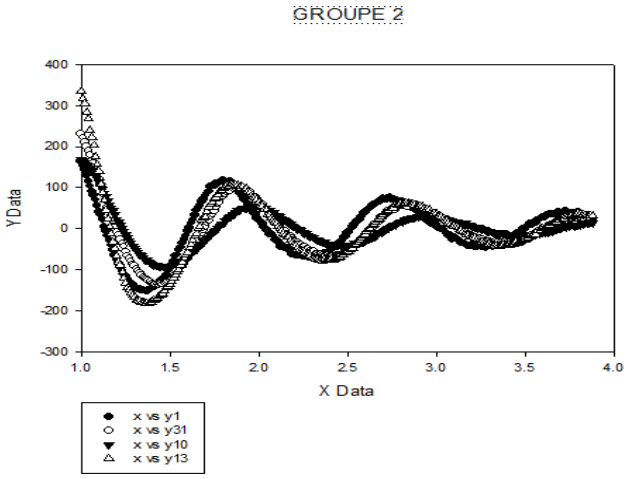
Pour vérifier cela, on veut montrer qu'on peut ajuster une triple des variables (b,c,d) unique pour chaque groupe.

1) Présentation de la méthode

Première étape, je superpose les données. Pour tous les échantillons dans même groupe, la valeur de x qui représente le temps est toujours de 1 à 4. Mais les valeurs initiales ne sont pas égales, par exemple, pour l'échantillon A1_4, x est de 1.0035 à 4.0035, pour l'échantillon A1_19, x est de 1.0036 à 4.0063 etc. Donc pour exprimer les données de façon cohérent, je crée un nouveau axe de x qui est de 1 à 3.88, la différence dans le modèle correspond bien à 15 minutes en réalité. Je crée un nouveau tableau avec le nouveau x et les valeurs de y de ces 4 échantillon du groupe. Et puis, je mets tous les points de ces 4 échantillons dans un même graphe.

Pour le groupe 1 et le groupe 2, on a les graphes suivants





Deuxième étape, il faut chercher un modèle commun pour chaque groupe. D'après la partie III, on peur voir que pour les deux groupes, le modèleest le plus utilisé (sauf les échantillons A2_1 et A1_4). Et on ne peut pas voir des différences remarquables entre les modèles complets et les modèles simplifiés. Donc dans cette partie, on va appliquer les modèles suivants pour les régressions généraux des ces deux groupes.

Groupe 1 : $y = f(x) = a * \exp(-x/d) * \sin\left(\frac{2*\pi*x}{b} + c\right)$

Groupe 2 : $y = f(x) = a * \exp(-x/d) * \sin\left(\frac{2*\pi*x}{b} + c\right)$

Je commence par la recherche d'une période (associée au paramètre b) commune au sein de la deuxième groupe. Je réalise deux régressions globales. Pour la première régression noté régression A, il n'y aura pas de paramètre à partager. Pour la deuxième régression noté régression B, le paramètre b sera partagé par tous les échantillons du groupe. Et puis on souhaite comparer les 2 modèles de régression en utilisant le test de fisher.

Soit n=4 le nombre d'échantillon de chaque groupe. Pour la régression A, on applique la fonction y=f(a, d, b, c) à la famille de données de chaque groupe où a=(a1, a2, a3, a4), d=(d1, d2, d3, d4) ,b=(b1, b2, b3, b4) et c=(c1, c2, c3, c4).

Pour la régression B, on applique la fonction y=f(a, d, b0, c) à la famille de données de chaque groupe où b0 est un paramètre partagé pour les membres de la famille, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4) et c=(c1, c2, c3, c4).

Pour la comparaison entre les deux modèles de régression, on utilise la statistique de test de fisher suivante :

$$F = \frac{\frac{SSshared - SSunshared}{DFshared - DFunshared}}{\frac{SSunshared}{DFunshared}}$$

où

SSshared = Somme des carrés du modèle qui a un ou des paramètres partagés (régression B)
SSunshared = Somme des carrés du modèle qui n'a pas de paramètre partagé (régression A)
DFshared = degré de liberté du modèle qui a un ou des paramètres partagés (régression B)
DFunshared = degré de liberté du modèle qui n'a pas de paramètre partagé (régression A)

.

Si on a trouvé que la différence de période entre les échantillons du groupe n'est pas significative, on conserve la valeur de la période noté b0. Je garde la régression B où a=(a1, a2, a3, a4), b=b0,d=(d1, d2, d3, d4) et c=(c1, c2, c3, c4). Pour tester la différence d'un autre paramètre du groupe par exemple la phase, on suppose que le régression B en applicant y=f(g, a, d, b0, c) est une autre régression A. Et puis, on applique la fonction y=f(a, d, b0, c0) à la famille de données de chaque groupe pour un autre régression B où b0 et c0 sont paramètres partagés, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4). On refait la comparaison pour le paramètre c avec la statistique F et on va comparer cette statistique calculé avec la valeur $F_{1-0.05,3,1138}$.

Si on a trouvé que la différence de période entre les échantillons du groupe est significative, on garde la régression A où a=(a1, a2, a3, a4), b=(b1, b2, b3, b4) ,d=(d1, d2, d3, d4) et c=(c1, c2, c3, c4). Pour tester la différence de phase du groupe on applique la fonction y=f(g, a, d, b, c0) à la famille de données de chaque groupe pour un autre régression B où b0 et c0 sont paramètres partagés, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4) ,b=(b1, b2, b3, b4). On refait la comparaison pour le paramètre c avec la statistique F et on va comparer cette statistique calculé avec la valeur $F_{1-0.05,3,1138} = 2.612716$.

On itère la même procedure pour les autres paramètres (l'amplitude a et l'amortissement d)

2) Les résultats d'analyse

Pour le premier groupe, on a les résultats de régression A suivants :

Analysis of Variance:

	DF	SS	MS
Regression	16	485307.9344	30331.7459
Residual	1140	14414.3847	12.6442
Total	1156	499722.3191	432.2857

Les estimations des paramètres de régression A sont suivants :

Paramètre	Estimation	erreur type	t_value	p_value
a1	219.6415	6.9440	31.6304	<0. 0001
a2	180.7391	5.6688	31.8829	<0. 0001
a3	31.0931	1.3294	23.3883	<0. 0001
a4	48.6930	2.5110	19.3922	<0. 0001
d1	1.2076	0.0269	44.8916	<0. 0001
d2	1.3048	0.0301	43.3033	<0. 0001
d3	1.3682	0.0441	31.0195	<0. 0001
d4	1.3298	0.0508	26.1814	<0. 0001
b1	0.9952	0.0027	362.3976	<0. 0001
b2	1.0164	0.0026	391.2372	<0. 0001
b3	1.0077	0.0033	303.8126	<0. 0001
b4	1.0016	0.0040	248.2834	<0. 0001
c1	1.9298	0.0288	67.0774	<0. 0001
c2	2.2388	0.0264	84.6586	<0. 0001
c3	2.5995	0.0344	75.6509	<0. 0001
c4	2.3378	0.0420	55.6180	<0. 0001

On a les résultats de régression B suivants :

Analysis of Variance:

	DF	SS	MS
Regression	13	484595.9578	37276.6121
Residuall	143	15126.3613	13.2339
Total	1156	499722.3191	432.2857

Les estimations des paramètres de régression B sont suivantes :

Paramètre	Estimation	erreur type	t_value	p_value

a1	222.9096	5.1122	43.6032	<0. 0001
a2	180.4919	4.5635	39.5513	<0. 0001
a3	31.1484	4.3736	7.1218	<0. 0001
a4	48.7851	4.5407	10.7441	<0. 0001
d1	1.1979	0.0193	62.1988	<0. 0001
d2	1.3012	0.0242	53.6606	<0. 0001
d3	1.3662	0.1445	9.4569	<0. 0001
d4	1.3285	0.0916	14.5072	<0. 0001
b0	1.0054	0.0014	711.4375	<0. 0001
c1	2.0298	0.0152	133.4947	<0. 0001
c2	2.1332	0.0157	135.9683	<0. 0001
c3	2.5772	0.0418	61.6156	<0. 0001
c4	2.3754	0.0295	80.6147	<0. 0001

SSunshared = 485307.9344 DFunshared = 16
SSshared = 484595.95 DFshared = 13

F=0.00782< $F_{1-0.05,3,1138}=2.612716$

On ne rejette donc pas l'hypothèse : Il n'existe aucune différence entre les deux ajustements pour chacune de nos données au sein d'un même groupe. Et b0 = 1.0054 = 25.13 heures. Il est donc possible d'ajuster une période unique pour le premier groupe.

Donc on garde la période commune pour tester l'existence de différence de la phase, on conserve la période commune et on suppose que le régression B en applicant y=f(a, d, b0, c) est une autre régression A :

Analysis of Variance:

	DF	SS	MS
Regression	13	484595.9578	37276.6121
Residuall	143	15126.3613	13.2339
Total	1156	499722.3191	432.2857

Et puis, on applique la fonction y=f(a, d, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0 et c0 sont paramètres partagés, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4).

Analysis of Variance:

	DF	SS	MS
Regression	10	479815.0306	47981.5031
Residuall	146	19907.2885	17.3711
Total	1156	499722.3191	432.2857

SSunshared = 484595.9578 DFunshared = 13
SSshared = 479815.0306 DFshared = 10

F=0.0428< $F_{1-0.05,3,1138}=2.612716$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein d'un même groupe. Et c0 = 2.0802 = 3120.3 heures. Il est donc possible d'ajuster une phase unique pour le premier groupe.

Donc on garde la période commune et la phase commune pour tester l'existence de différence de l'amortissement, on conserve la période commune et phase commune on suppose que le régression B en applicant y=f(a, d, b0, c0) est une autre régression A :

Analysis of Variance:

	DF	SS	MS
Regression	10	479815.0306	47981.5031
Residuall	146	19907.2885	17.3711
Total	1156	499722.3191	432.2857

Et puis, on applique la fonction y=f(g, a, d0, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0, c0 et d0 sont paramètres partagés, a=(a1, a2, a3, a4) :

Analysis of Variance:

	DF	SS	MS
Regression	7	479411.1459	68487.3066
Residuall	149	20311.1732	17.6773
Total	1156	499722.3191	432.2857

SSunshared = 479815.0306 DFunshared = 10
SSshared = 479411.1459 DFshared = 7

F=0.00084< $F_{1-0.05,3,1138}=2.612716$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein d'un même groupe. Et d0 = 1.2507. Il est donc possible d'ajuster les données avec un coefficient d'amortissement unique pour le premier groupe.

Pour le deuxième groupe, on a les résultats de régression A suivants :

Analysis of Variance:

	DF	SS	MS
Regression	16	5304594.7666	331537.1729
Residuall	140	112090.7928	98.3253
Total	1156	5416685.5594	4685.7142

On a les résultats de régression B suivants :

Analysis of Variance:

	DF	SS	MS
Regression	13	5279030.8570	406079.2967
Residuall	143	137654.7024	120.4328
Total	1156	5416685.5594	4685.7142

Les estimations des paramètres de régression B sont dans l'annexe.

SSunshared = 5304594.7666 DFunshared = 16
SSshared = 5279030.8570 DFshared = 13

F=0.0257< $F_{1-0.05,3,1138}=2.612716$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux ajustements pour chacune de nos données au sein d'un même groupe. Et b0 = 0.9531 = 23.83 heures. Il est donc possible d'ajuster les données avec une période unique pour le premier groupe.

Donc on garde la période commume pour tester l'existence de différence de la phase, on conserve la période commune et on suppose que le régression B en applicant y=f(a, d, b0, c) est une autre régression A :

Analysis of Variance:

	DF	SS	MS
Regression	13	5279030.8570	406079.2967
Residuall	143	137654.7024	120.4328
Total	1156	5416685.5594	4685.7142

Et puis, on applique la fonction y=f(a, d, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0 et c0 sont paramètres partagés, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4).

Analysis of Variance:

	DF	SS	MS
Regression	10	4917455.2531	491745.5253
Residuall	146	499230.3063	435.6285
Total	1156	5416685.5594	4685.7142

SSunshared = 5279030.8570 DFunshared = 13
SSshared = 4917455.2531 DFshared = 10

F=0.2968< $F_{1-0.05,3,1138}=2.612716$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour

chacune de nos données au sein d'un même groupe. Et c0 = 1.5757 = 2363.55 heures. l est donc possible d'ajuster les données avec une phase unique pour le premier groupe.

Donc on garde la période commune et la phase commune pour tester l'existence de différence de l'amortissement, on conserve la période commune et phase commune on suppose que le régression B en applicant y=f(a, d, b0, c0) est une autre régression A :

Analysis of Variance:

	DF	SS	MS
Regression	10	4917455.2531	491745.5253
Residuall	146	499230.3063	435.6285
Total	1156	5416685.5594	4685.7142

Et puis, on applique la fonction y=f(g, a, d0, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0, c0 et d0 sont paramètres partagés, a=(a1, a2, a3, a4) :

Analysis of Variance:

	DF	SS	MS
Regression	7	4864632.2931	694947.4704
Residuall	149	552053.2662	480.4641
Total	1156	5416685.5594	4685.7142

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein d'un même groupe. Et d0 = 1.1254. Il est donc possible d'ajuster les données avec un coefficient d'amortissement unique pour le premier groupe.

<D'après les résultats, on peut conclure qu'au sein d'une équipe des animaux, l'horloge biologique admet une oscillation assez proche. Pour chaque groupe, on peut l'ajuster avec une seule période, une seule phase et un seul coefficient d'amortissement. L'existence des caractéristiques entre les différents échantillons d'un même groupe est évidente.

IX. Comparaison les caractéristiques entre les groupes

Pour mieux comprendre l'influence de nourriture grasse sur l'horloge biologique au sein du cerveau, on va réaliser une comparaison entre les deux groupes en comparant 3 quantités qu'on a déjà étudiés dans la partie précédente. Elles sont la période (correspondant au paramètre <b>), la phase (correspondant au paramètre <c>) et le coefficient d'amortissement (correspondant au paramètre <d>).

On aimerait réaliser ces comparaisons par la méthode ANOVA, mais il s'avère que tous les quantités qu'on a étudiées sont obtenues avec des erreurs types. Il faudrait donc montrer que les erreurs sont négligeables pour chaque groupe, afin de pouvoir considérer que les valeurs de la période, la phase, et le coefficient d'amortissement obtenues sont les valeurs exactes.

Pour la comparaison, on peut aussi par la méthode de la régression globale. On va comparer le résultat de cette méthode avec celui de la méthode ANOVA à la fin.

Donc, il y a deux objectives concrètes dans cette partie est :

1) De vérifier que si la variance intra-groupe (ici, plutôt intra-échantillon) est négligeable par rapport à la variance inter-groupe (ici, plutôt inter-échantillon)

2) De comparer les 3 quantités entre les 2 groupes par la méthode ANOVA ou la méthode du régression global (général)

1. Étude de la variance intra-échantillon et la variance inter- échantillon

D'après le cour de la régression, on sait que le résultat d'ANOVA :

$$SC_{total} = SC_{inter-échantillon} + SC_{intra-échantillon}$$

$$ddl_{Intra} = ddl_{Total} - ddl_{Inter} = (N-1) - (P-1)$$

ddl= degré de liberté SC = Somme des Carrés

n = Nombre total des observations p = Nombre de échantillons au sein de groupe

Pour vérifier les erreurs intra-échantillon sont négligeables pour chaque groupe, ici on va utiliser encore un test de fisher. On calcule la statistique du test

$$F = \frac{SC_{inter-échantillon} \div DDL_{inter}}{SC_{intra-échantillon} \div DDL_{intra}}$$

Si F > $F_{p,n-p-1}$, on rejette H0, et on peut dire qu'il y a une différence significative entre les groupes, les erreurs intra-échantillon peuvent être négligeable par rapport aux erreurs inter-échantillon.

Si F < $F_{p,n-p-1}$, on rejette H1, et on peut dire qu'il n'y a pas de différence significative entre les groupes, les erreurs intra-échantillon ne peuvent pas être négligeable par rapport aux erreurs inter-échantillon.

On peut voir les résultats d'ANOVA pour les deux groupes :

Groupe 1

	DDL	SC	MSC	Valeur F	P value
Inter_échantillon	3	358	119.2	0.276	0.843
Intra_échantillon	1152	497535	431.9		
Total	1155	497893	431.0762		

Groupe 2

	DDL	SC	MSC	Valeur F	P value
Inter_échantillon	3	1867	622	0.133	0.941
Intra_échantillon	1152	5409979	4696		
Total	1155	5411846	4697.783		

Les codes en R sont dans l'annexe. Les p-valeurs sont supérieures à 0.05, les tests ne sont pas significatifs. Il n'y a pas de différence significative entre les groupes, les erreurs intra-échantillon ne peuvent pas être négligeables par rapport aux erreurs inter-échantillon. On ne peut pas considérer que les valeurs de la période, la phase, et le coefficient d'amortissement obtenues sont les valeurs exactes. Donc on n'a pas droit d'utiliser la méthode ANOVA directement pour la comparaison des paramètres entre les groupes. Mais on a encore la méthode de la régression générale.

2. Comparaison avec les données originales en utilisant la régression générale

Pour résoudre le problème des erreurs standards, on va comparer les données originales entre les deux groupes en utilisant la méthode de régression générale et même idée de la partie VII. Dans la partie VII, pour exprimer les données de façon cohérent, je crée un nouveau axe de x qui est de 1 à 3.88, la différence $x_{i+1}-x_i=0.01$ dans le modèle correspond bien à 15 minutes en réalité. On va encore utiliser cet axe de x dans cette partie. J'ai crée un nouveau tableau avec le nouveau x et les valeurs de y de tous les 8 échantillons de l'ensemble de ces 2 groupes.

Pour la réalisation de régressions générales, on a trouvé que la modèle simplifié

$$y_i=f(x_i)=a*\exp(\frac{-x_i}{d})*\sin(\frac{2*\pi*x_i}{b}+c)$$

est le meilleur modèle pour les 8 échantillons. En outre, d'après les résultats des tests de fisher, on savait déjà qu'au sein d'une équipe des animaux, l'horloge biologique admet une oscillation assez proche. Pour chaque groupe, on peut ajuster une seule période, une seule phase et un seul coefficient d'amortissement.

Donc dans cette partie, on va encore utiliser le même modèle simplifié pour la régression générale. Ainsi, si à l'issue du test qu'on conclut à la significativité (on rejette H0), on saura qu'il y a une

différence significative entre les périodes des deux groupes comparés.

En prenant le même principe de la méthode dans VII, Je réalise deux régressions générales. Pour la première régression noté régression A, il n'y aura pas de paramètre à partager. Pour la deuxième régression noté régression B, le paramètre qui me concerne sera partagé par les 8 échantillons. Et puis on souhaite comparer les 2 modèles de régression en utilisant le test de fisher en utilisant la statistique :

$$F = \frac{\frac{SSshared - SSunshared}{DFshared - DFunshared}}{\frac{SSunshared}{DFunshared}}$$

Premièrement, on va comparer la période entre les groupes.

On a obtenu le résultat de régression A suivant :

Analysis of Variance:

	DF	SS	MS
Regression	32	5789902.7009	180934.4594
Residual	2280	126505.1775	55.4847
Total	2312	5916407.8784	2558.9999

On a obtenu le résultat de régression B en partageant le paramètre b suivant :

Analysis of Variance:

	DF	SS	MS
Regression	25	5746933.6637	229877.3465
Residual	2287	169474.2148	74.1033
Total	2312	5916407.8784	2558.9999

Les estimations des paramètres de régression A et régression B sont dans l'annexe.

SSunshared = 5789902.7009 DFunshared = 32
SSshared = 5746933.6637 DFshared = 25

F=0.0339< $F_{1-0.05,7,2278}$=2.013594

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein de ces 8 échantillons. Donc on peut conclure pour la période, il n'y a pas de différences significatives entre les groupes.

Donc on garde la période commume pour tester l'existence de différence de la phase, on conserve la période commune et on suppose que le régression B en applicant y=f(a, d, b0, c) est une autre

régression A :

Analysis of Variance:

	DF	SS	MS
Regression	25	5746933.6637	229877.3465
Residual	2287	169474.2148	74.1033
Total	2312	5916407.8784	2558.9999

Et puis, on applique la fonction y=f(a, d, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0 et c0 sont paramètres partagés, a=(a1, a2, a3, a4), d=(d1, d2, d3, d4).

Analysis of Variance:

	DF	SS	MS
Regression	18	5371557.9329	298419.8852
Residual	2294	544849.9456	237.5109
Total	2312	5916407.8784	2558.9999

SSunshared = 5746933.6637 DFunshared = 25
SSshared = 5371557.9329 DFshared = 18

F=0.233< $F_{1-0.05,7,2278}=2.013594$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein de ces 8 échantillons. Donc on peut conclure pour la phase, il n'y a pas de différences significatives entre les groupes.

Donc on garde la période commune et la phase commune pour tester l'existence de différence de l'amortissement, on conserve la période commune et phase commune on suppose que le régression B en applicant y=f(a, d, b0, c0) est une autre régression A :

Analysis of Variance:

	DF	SS	MS
Regression	18	5371557.9329	298419.8852
Residual	2294	544849.9456	237.5109
Total	2312	5916407.8784	2558.9999

Et puis, on applique la fonction y=f(a, d0, b0, c0) à la famille de données du premier groupe pour un autre régression B où b0, c0 et d0 sont paramètres partagés, a=(a1, a2, a3, a4) :

Analysis of Variance:

	DF	SS	MS
Regression	11	5324131.7480	484011.9771
Residual	2301	592276.1304	257.3994
Total	2312	5916407.8784	2558.9999

SSunshared = 5371557.9329 DFunshared = 18
SSshared = 5324131.7480 DFshared = 11

F=0.0227< $F_{1-0.05,7,2278} = 2.013594$

On ne rejette donc pas l'hypothèse: Il n'existe aucune différence entre les deux régressions pour chacune de nos données au sein de ces 8 échantillons. Donc on peut conclure pour le coefficient d'amortissement, il n'y a pas de différences significatives entre les groupes.

X. Conclusion

D'après les résultats de comparaison, on constate que l'oscillation de activité dans le cerveau pour les deux groupe d'animaux admet quasiment le comportement identique de la période, la phase et le coefficient d'amortissement, bien que leur amplitudes varient intuitivement. Donc d'après le résultat statistique, la modification de nourriture ne se produit aucune influence sur le comportement d'oscillation dans le cerveau d'animal expérimental au niveau de la période, la phase et l'amortissement.

Donc, d'après les résultats statistique, l'augmentation d'ingestion de nourriture grasse pendant la période d'endormi ne peut pas impliquer la relation fonctionnelle entre l'horloge circadienne et le mécanisme hédonistique central.

D'autre point de vue, les données que l'expérience se produit ne peut pas présenter significativement de différence sur le comportement d'oscillation dans le cerveau des animaux expérimentaux, car le nombre des échantillons effectifs n'est pas assez important. Donc finalement, je doit conseiller aux chercheurs de répéter l'expérience avec davantage d'animaux, et ils ont une chance de trouver un changement du comportement d'oscillation au niveau de la période, la phase et l'amortissement.