Méthodes d'analyse de données en régression non linéaire

Katell Mellac

Sommaire

Remerciements……..p.2

I) Introduction……...p.3

1°) Neurobiologie des rythmes……..p3

a) Présentation de l'institut des neurosciences cellulaires et intégratives……p3

b) Présentation du sujet de stage…… p4

2°) La régression non linéaire et l'algorithme de Gauss-Newton……..p7

a) La régression non linéaire…….....p7
b) Estimation de paramètres……..p7
c) Algorithme de Gauss-Newton…….p8

II) Ajustements de nos données par des régressions non-linéaires et validations des modèles……....p9

1°) Application à un échantillon……...p9

a) Vérification de la normalité et de l'homoscédasticité des résidus……p10
b) Estimation des paramètres……p12
c) Matrice de corrélations entre les paramètres……p14

III) Elimination des paramètres responsables de la multicolinéarité ……...p16

IV) Calcul des nouvelles périodes ……...p19

V) Recherche d'une période commune au sein de chaque groupe……p22

1°) Présentation de la méthode……...p22

2°) Recherche d'une période commune pour les échantillons du groupe CoGC+INL……p23

VI) Comparaison entre les groupes……p29

1°) Avec une régression générale……p31

2°) Avec la fonction « aov » de R……p33

3°) Présentation d'une méthode permettant de négliger les erreurs associées à la période……p35

4°) Résultats des ANOVA……...p40

VII) Conclusion……...p41

ANNEXE

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, ainsi que pour m'avoir permis d'effectuer mon stage au sein de l'INCI.

Je remercie également **Catherine Jaeger**, doctorante à l'INCI, pour toutes ses explications concernant son sujet de thèse, les protocoles expérimentaux, pour m'avoir permis d'assister aux manipulations au sein des laboratoires, ainsi que pour l'aide qu'elle m'a apportée tout au long de mon stage et plus généralement concernant le fonctionnement de l'horloge biologique et de la rétine. Je la remercie également pour son accueil chaleureux, sa patience, sa gentillesse et son soutien.

Je remercie également tous les autres doctorants pour leur accueil, leur gentillesse et pour m'avoir aidée à mieux appréhender leur travail de biologiste.

Pour finir, je remercie toutes les personnes du laboratoire de neurobiologie des fonctions rythmiques pour leur accueil et leur bonne humeur.

I) introduction

1°) La neurobiologie des rythmes

a)Présentation de l'institut des neurosciences cellulaires et intégratives.

Les recherches en neurosciences sont devenues une priorité actuellement, étant donné le nombre croissant de patients atteints de maladies du cerveau, ceci étant lié notamment à l'augmentation de l'espérance de vie.

L'INCI, l'institut des neurosciences cellulaires et intégratives, est un institut de recherche qui se concentre sur trois thèmes majeurs, qui sont les rythmes biologiques, la neurosécrétion et la nociception.

L'institut se divise ainsi en trois départements :

⇨ **Neurobiologie des fonctions rythmiques**

⇨ **Physiologie des réseaux neuronaux**

⇨ **Nociception et douleur**

La nociception est une fonction défensive, permettant l'intégration au niveau du système nerveux central d'un stimulus douloureux via l'activation des nocicepteurs (récepteurs à la douleur) cutanés, musculaires et articulaires.

La physiologie des réseaux neuronaux s'intéresse aux phénomènes de libération par les neurones de molécules informatives destinées à la communication avec d'autres cellules.

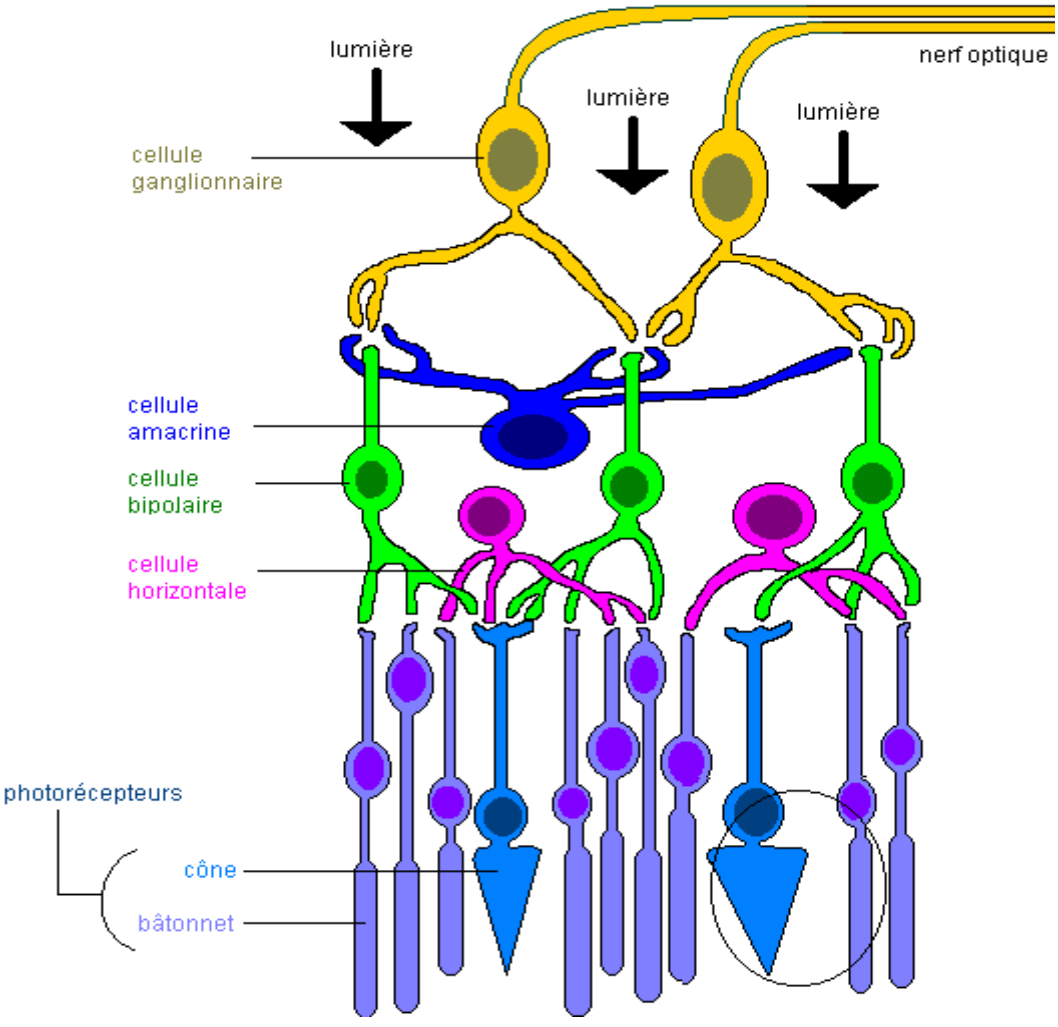
Pour mon stage, j'ai intégré le département de neurobiologie des rythmes. Au sein de ce département, quatre équipes cherchent à comprendre les mécanismes nerveux et endocriniens impliqués dans le contrôle des rythmes biologiques. Ces rythmes permettent à l'organisme de s'adapter aux variations journalières et saisonnières de l'environnement. Les recherches sont effectuées sur des rats, des souris et des hamsters.

L'INCI est une unité propre du CNRS et est également partenaire de l'université de Strasbourg

b) Présentation du sujet de stage

Lors de ce stage, l'étude statistique a porté sur des données représentant l'activité rythmique de rétines de souris. La rétine est un tissu sensoriel composé de trois couches de neurones connectées par des couches de contacts synaptiques :

La couche la plus externe (appelée photoreceptor layer ou **PRL**), qui tapisse le fond du globe oculaire, contient les neurones photosensibles appelés photorécepteurs. La couche médiane, appelée inner nuclear layer ou **INL,** contient différents types de neurones qui reçoivent et traitent les informations photiques transmises par les photorécepteurs. Enfin, la couche la plus interne au globe oculaire, appelée ganglion cell layer ou **GCL,** transmet les informations reçues de l'INL au cerveau via le au nerf optique.



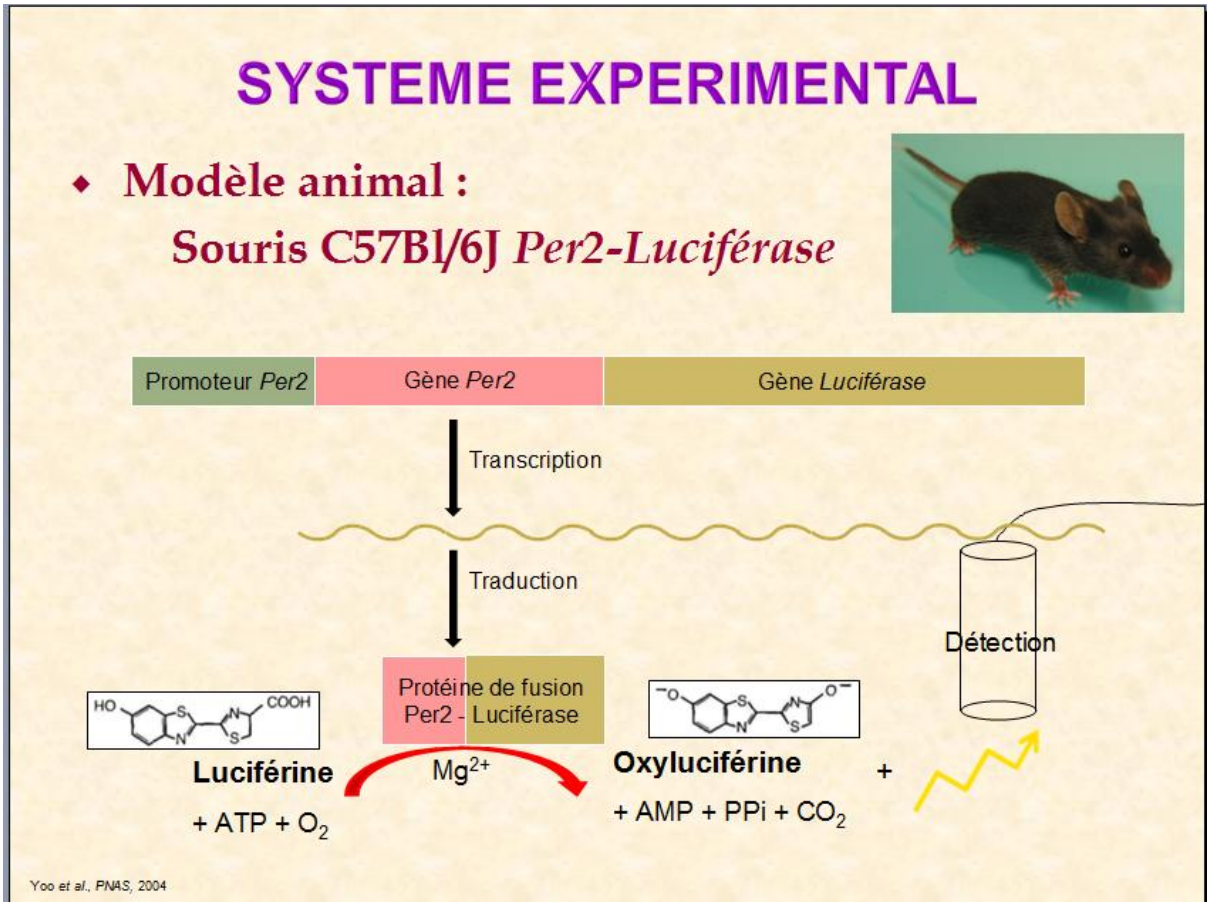
SCHEMA DE LA RETINE AVEC DETAIL DES TROIS COUCHES

Il a déjà été démontré que la rétine contenait une horloge biologique interne, c'est-à-dire un ensemble de gènes qui s'expriment en boucle de manière autonome, induisant sur une période de 24h environ leur propre expression oscillatoire ainsi que celle de nombreux autres gènes qui contrôlent la physiologie de la rétine. Le travail de thèse de C. Jaeger consiste à répondre à plusieurs questions concernant cette horloge locale :

-L'horloge biologique de la rétine se situe-t-elle dans une couche de la rétine ou dans toutes les couches ?

- Comment les différentes couches communiquent-elles entre-elles pour se transmettre les informations rythmiques ?

Le modèle biologique utilisé est une souris transgénique créée pour exprimer un rapporteur visible du fonctionnement de l'horloge biologique : la luciférase, présente dans les cellules de l'animal suite à la modification génétique, catalyse une réaction chimique bioluminescente et ainsi, les variations de la bioluminescence émise par les cellules reflètent les variations oscillatoires de l'un des gènes de l'horloge biologique, appelé Per2.



La lignée de souris est une lignée dire « knock-in ». La transgenèse est insérée dans le génome de la souris à la place du gène endogène, sur les deux allèles. Dans ce cas c'est une transgenèse de fusion, et donc les variations de la bioluminescence enregistrée reflètent à la fois la régulation de l'expression du gène Per2 et la vie de la protéine PER2

Pour répondre aux questions posées, différentes préparations de rétines ont été placées en culture dans un appareil de mesure de la bioluminescence et ont été suivies sur plusieurs jours afin de vérifier la présence d'une activité oscillatoire et, le cas échéant, de mesurer la période des oscillations détectées. Les différentes préparations rétiniennes sont :

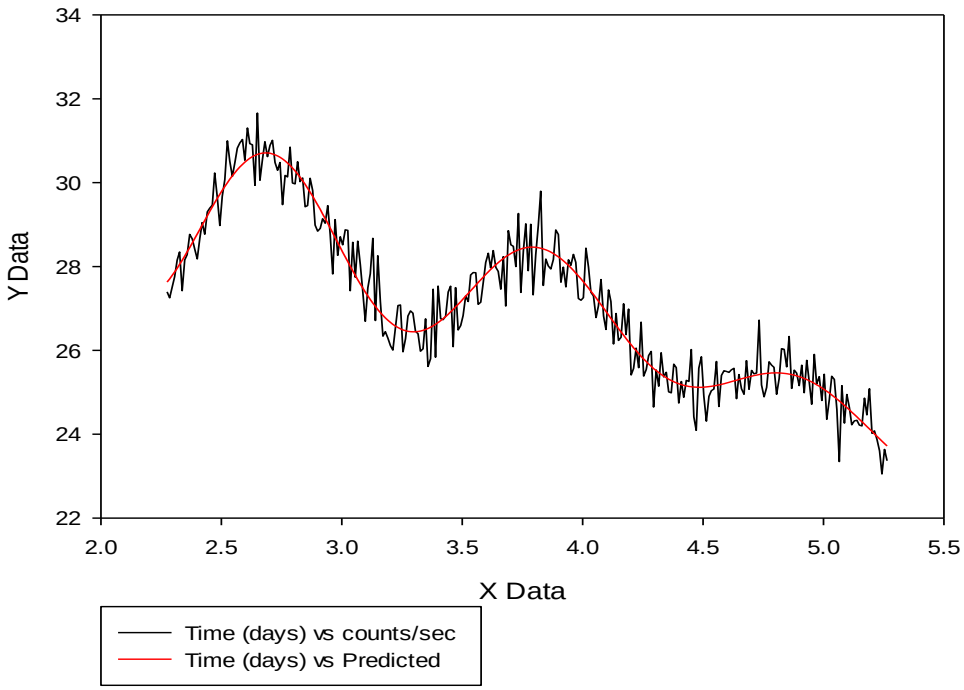
- « **Whole retina** » : rétine entière
- Chacune des trois couches de la rétine isolée des autres : **PRL, INL, GCL**
- Les couches **GCL** et **INL** (**GCL+INL**) après avoir ôté la couche **PRL** de la rétine, et les couches **PRL** et **INL** (**PRL+INL**) après avoir ôté la couche **GCL**
- L'ensemble des cellules de la rétine dissociées les unes des autres (cultivées ensemble mais sans contacts physiques) (**Dissociated cells**)
- Des co-cultures de **GCL** et **INL** (**CoGCL+INL**), c'est-à dire des couches **GCL** et **INL** séparées mais cultivées ensemble.

On obtient ainsi huit groupes contenant chacun un nombre différent d'échantillons, pour un total de 120 échantillons.

Tous les types de préparations présentent une activité oscillatoire.

La première oscillation enregistrée n'est pas prise en compte pour l'analyse car elle n'est pas représentative de la réalité : il faut un peu de temps aux cellules en culture pour s'adapter aux conditions de culture. Les courbes sont donc analysées sur trois jours à partir du temps t=0 au pied de la seconde oscillation.

Les périodes sont calculées par l'ajustement d'une régression non linéaire sur les données brutes (activité oscillatoire en fonction de temps).



Enregistrement de l'oscillation des cellules(en noir) et courbe ajustée(en rouge). La courbe conservée commence au pied de la seconde oscillation

Pour calculer les périodes, on est parti des données brutes (oscillation en fonction de temps) et on les a ajustées à l'aide d'une régression non linéaire.

2°) La régression non linéaire et l'algorithme de Gauss-Newton

a) La régression non linéaire

La régression non linéaire a pour but d'ajuster un modèle non linéaire pour un ensemble de valeurs afin de déterminer la courbe qui se rapproche le plus de celle des données de Y en fonction de x.

Le modèle de régression linéaire s'écrit $Y_i=f(x_i,\Theta)+\varepsilon_i$

I=1,…..,n

-La loi de probabilité sur ε_i est une loi normale, centrée réduite et de variance σ^2_i finie.

-Les ε_i sont indépendants entre eux.

-Y_i représente l'observation i de la variable dépendante

-Θ représente un vecteur à p composantes de paramètres généralement inconnus.

-La fonction f est la fonction de régression, la plupart du temps non linéaire. Elle dépend d'une variable réelle x et de paramètres Θ.

b) Estimation des paramètres

Comme en régression linéaire, les paramètres d'un modèle de régression non linéaire sont estimés en minimisant la somme des carrés des résidus du modèle.
C'est-à-dire qu'on cherche à minimiser l'expression suivante :

$SSR(\Theta)=\sum_{i=1}^{n}(Y_i - f(x_i,\Theta))^2$

Pour cela, il faut dériver cette somme par rapport à chacun de ses paramètres et chercher les solutions qui annulent les dérivés.
On peut réécrire ceci sous forme vectorielle :
SSR(Θ)=(Y- f(x,Θ))(Y- f(x,Θ))'
SSR(Θ)=YY'-2Y f(x,Θ)'+f(x,Θ) f(x,Θ)'

On dérive cette expression par rapport à toutes les composantes du vecteur Θ à p paramètres, et on annule toutes les dérivées partielles.
On obtient les conditions du premier ordre qui doivent être vérifiées pour toute estimation du vecteur Θ qui correspond à un minimum intérieur de SSR(Θ) . Ces conditions du premier ordre, ou équations normales, sont :

-2F(x, $\hat{\Theta}$)'Y+2F(x, $\hat{\Theta}$)'f(x, Θ)=0 (I)

Où la matrice F(x, Θ) de dimension n×p est composée d'éléments du type :

$F_{i,j}(x, \Theta) \equiv \frac{\partial f_{j(x\Theta)}}{\partial_{\Theta j}}$

Le fait que chaque vecteur de l'équation (I) possède p éléments implique l'existence de p équations normales déterminant les p composants de Θ.
Finalement, on obtient des équations qu'on ne peut la plupart du temps pas résoudre de manière analytique.
Cependant, il existe des algorithmes qui permettent d'estimer les paramètres. Nous nous intéresserons ici à l'algorithme de Gauss-Newton, qui est utilisé par défaut par la fonction « nls » de R.

c) Algorithme de Gauss-Newton

L'algorithme est basé sur un développement en série de Taylor au voisinage des valeurs initiales des paramètres, qu'on notera Θ^0.

Dans notre étude, on sera amenés à estimer 8 paramètres.

$f(x_i, \Theta)=f(x_i, \Theta^0) + d_{1,i}(\Theta_1 - \Theta_1^0)+ d_{2,i}(\Theta_2 - \Theta_2^0)+ d_{3,i}(\Theta_3 - \Theta_3^0)+ d_{4,i}(\Theta_4 - \Theta_4^0)$

$+ d_{5,i}(\Theta_5 - \Theta_5^0)+ d_{6,i}(\Theta_6 - \Theta_6^0)+ d_{7,i}(\Theta_7 - \Theta_7^0)+ d_{8,i}(\Theta_8 - \Theta_8^0)$

Avec $d_{a,i}=\frac{\partial \partial_{f(xi,\Theta)}}{\partial_{\Theta_a}} \mid_{\Theta=\Theta^0}$

Avec a=1,…,8

Ce qui donne, en écriture matricielle :

$\eta(\Theta)= \eta(\Theta^0) + V^0(\Theta - \Theta^0)$

Où $\eta(\Theta)$=f(x, Θ), $\eta(\Theta^0)$ =f(x, Θ^0), V^0 est la matrice n× 8 des dérivées partielles en $\Theta = \Theta^0$

Et x=$(x_1, x_2, \ldots, x_n)$

Pour obtenir les premiers estimateurs, il faut calculer :

$b_0 = [(V^0)'(V^0)]^{-1}(V^0)'[Y- \eta(\Theta^0)]$

On résout b_0 et on prend comme nouvel estimateur $\Theta^1=b_0 +\Theta^0$.

On calcule SSR(Θ^1). Si SSR(Θ^1)< SSR(Θ^0) alors on peut dire que Θ^1 est un meilleur estimateur que Θ^0. Si tel est le cas, on réitère ce procédé en remplaçant Θ^0 par Θ^1 (et V^0 par V^1) et on obtient une nouvelle série d'estimateur.

On calcule b_1. A partir de là, on peut calculer $\Theta^2 = b_1 + \Theta^1$.

On continue cette procédure jusqu'à la convergence.

II) Ajustements de nos données par des régressions non-linéaires et validations des modèles.

Tous les tests ont été réalisés au seuil $\alpha = 0.05$

On a travaillé sur l'équation suivante :

Y_i =c+(d*x_i)+(e*(x_i^2))+(g*(x_i^3))+(a-b*x_i)*sin(2*pi*(x_i+phi)/tau)+ε_i

où les Y_i représentent l'activité de la rétine et les x_i représentent le temps, exprimé en heures.

Phi représente la phase et tau représente la période. Les autres paramètres n'ont pas de signification biologique. **c, d, e et g** sont les paramètres de la ligne de base. Le paramètre tau est celui dont l'estimation nous intéresse le plus, car l'étude biologique porte sur la période des rétines.

Pour chaque échantillon, on a procédé à un ajustement de la courbe par une régression non linéaire afin d'obtenir une estimation du paramètre qui nous intéresse, le paramètre tau .

Le première partie du stage a consisté en la validation des régressions effectuées précédemment par Catherine. Il s'agissait de vérifier les conditions de normalité de d'homoscédasticité des résidus. Le travail de régression avait été effectué sous Sigmaplot. Ce logiciel nous donne les valeurs estimées, la courbe des observations en fonction du temps et son ajustement, ainsi que les valeurs des VIF (les facteurs d'inflation de la variance)

Les vérifications ont été faites avec R, une partie du code a été mise en annexe.

Comme on a effectué 120 régressions, on ne rendra compte ici que d'une seule vérification, les autres s'étant toutes faites de la même manière.

1°)Application à un échantillon

Voici un exemple pour un échantillon PR (couche des photorécepteurs)

On a vérifié que les résidus étaient distribués selon une loi normale par un test de Shapiro-Wilk. Pour vérifier l'homoscédasticité des résidus, on a utilisé un test de Bartlett.

On a également fait apparaitre les matrices de corrélations afin d'identifier les paramètres qui étaient liés entre eux.

On a effectué des tests de Durbin-Watson pour vérifier l'indépendance entre les données, cependant ces derniers aboutissaient parfois à la conclusion qu'il y avait un lien entre différents enregistrements. Il aurait fallu supprimer des données pour que les tests donnent des résultats satisfaisants. Or, cela aurait induit une perte de données inacceptable pour les biologistes. On n'en a donc pas tenu compte, bien qu'en toute rigueur il faille vérifier l'indépendance.

On va prendre l'exemple de l'échantillon **21A_PR3** appartenant au groupe PRL (cellules photo récepteur)

On a effectué un ajustement à l'aide de la fonction « nls » de R (non linear square).

a) Vérification de la normalité et de l'homoscédasticité des résidus.

On récupère les résidus du modèle à l'aide de la fonction « residuals » de R.

Les résidus s'expriment par $\widehat{eij} = Y_{ij} - \overline{Y_{i.}}$ **avec** $\overline{Yi.} = \sum_{J=1}^{J} Y_{ij}$

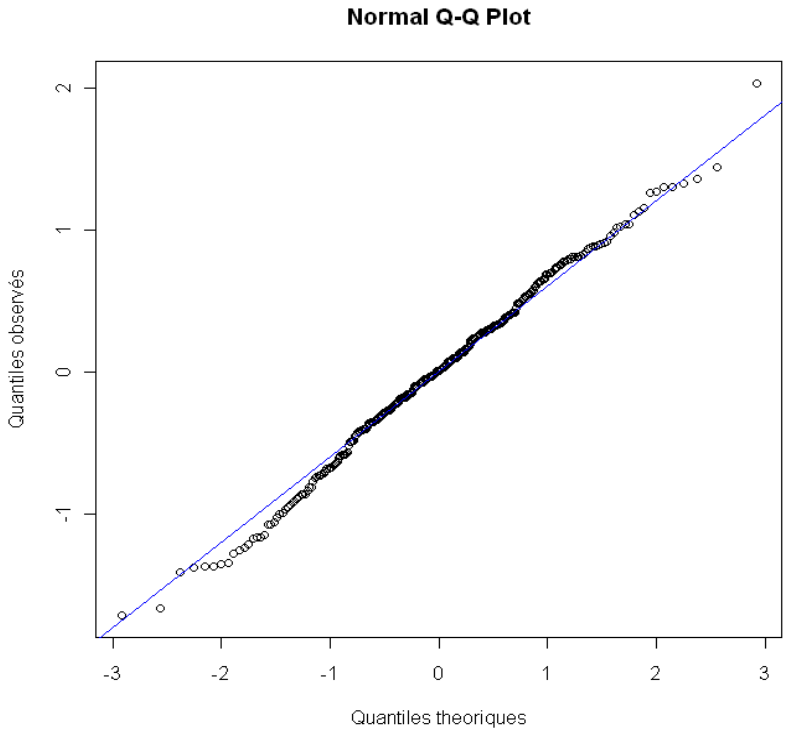
residus<-residuals(fit)

On effectue un test de Shapiro-Wilk sur les résidus :

shapiro.test(residus)

```
Shapiro-Wilk normality test

data:  residus
W = 0.9955, p-value = 0.567
```

On trace également la droite de Henry :



La p-value du test de Shapiro-Wilk montre qu'on peut conserver l'hypothèse H_0 de normalité des résidus.

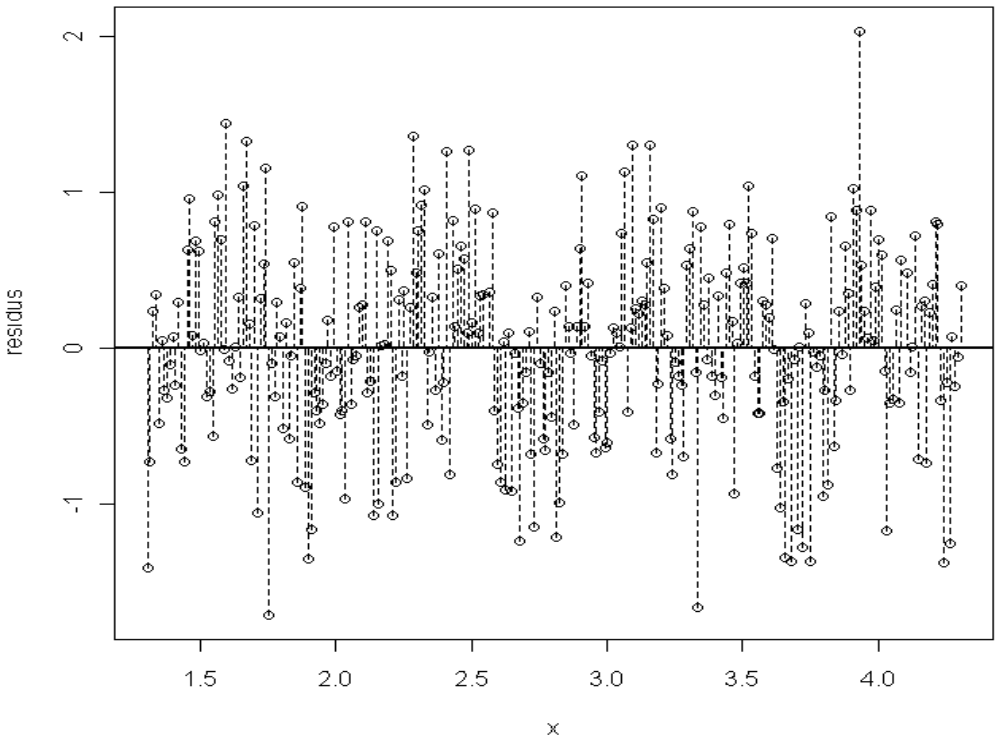
Le tracé de la droite de Henry nous permet d'aboutir à la même conclusion. La distribution observée est en effet en adéquation avec une distribution normale.

-Vérifions maintenant l'homogénéité des résidus à l'aide d'un test de Bartlett

bartlett.test(list(residus,x))

```
    Bartlett test of homogeneity of variances


data:  list(residus, x)

Bartlett's K-squared = 24.9368, df = 1, p-value = 5.924e-07
```

Traçons le graphique de répartition des résidus.



On constate que ceux-ci ne sont pas uniformément répartis autour d'une ligne centrale. Le test de Bartlett et le graphique démontrent qu'on doit rejeter l'hypothèse H_0 d'homogénéité des résidus.

L'exemple 21A_PR3 n'est pas une exception. La plupart des vérifications des régressions faites sur les échantillons donneront le même résultat : on rejette l'hypothèse d'homogénéité des résidus.

Cependant, l'hypothèse de normalité est toujours conservée. Les conclusions du test de Bartlett pour chaque échantillon sont données en annexe.

On a tenté de résoudre ce problème en appliquant des transformations sur la variable dépendante. On lui a appliqué plusieurs fonctions (logarithme, racine, puissance), mais on n'a pas réussi à obtenir l'homogénéité des résidus.

On a donc choisi de se baser sur d'autres critères pour valider les régressions. On a regardé graphiquement la qualité de l'ajustement du modèle aux données.

b) Estimation des paramètres

Pour l'échantillon **21A_PR3**, on a l'ajustement suivant :

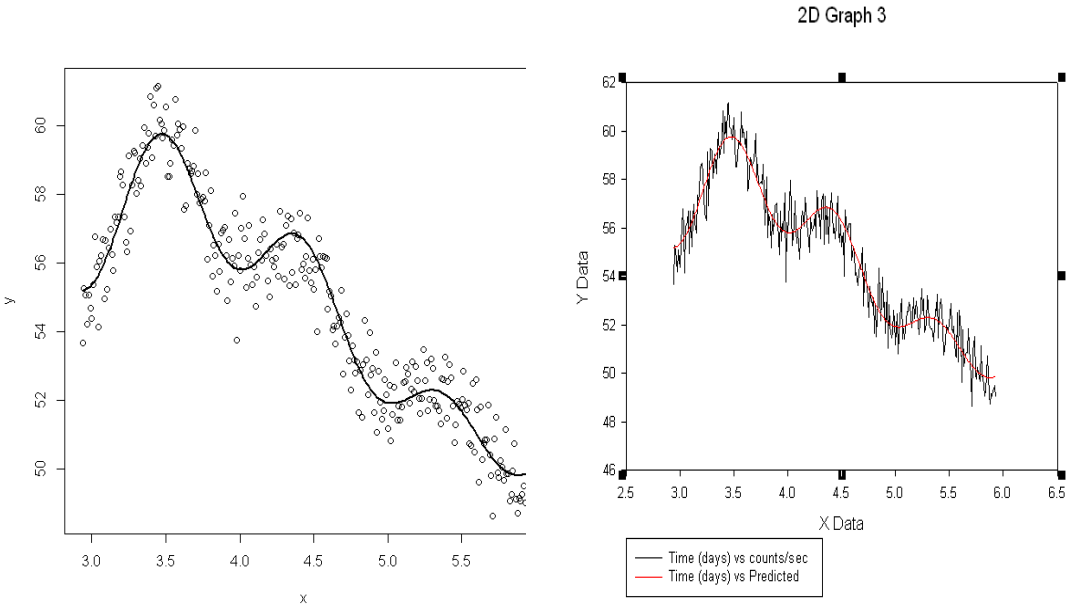
```
Formula: y ~ c + (d * x) + (e * (x^2)) + (g * (x^3)) + (a - b * x) * sin(2 *
    pi * (x + phi)/tau)


Parameters:
     Estimate Std. Error t value Pr(>|t|)
c   -33.22479    9.45662  -3.513 0.000516 ***
d    65.60647    6.62606   9.901  < 2e-16 ***
e   -15.03807    1.51763  -9.909  < 2e-16 ***
g     1.07166    0.11375   9.421  < 2e-16 ***
a     3.02892    0.40557   7.468 1.04e-12 ***
b     0.40340    0.09004   4.480 1.09e-05 ***
phi   0.61467    0.05747  10.696  < 2e-16 ***
tau   0.96343    0.01169  82.385  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.7917 on 280 degrees of freedom


Number of iterations to convergence: 7
Achieved convergence tolerance: 3.118e-06
```

Tous les paramètres sont significativement différents de zéro au seuil $\alpha = 0.05$

On obtient donc l'équation suivante :

$$Y_i = -33.22479 + 65.60647 \times x_i + -15.03807 \times x_i^2 + 1.07166 \times x_i^3 + (3.02892 - 0.40340 \times x_i) \times \sin(2 \times pi \times (x_i + 0.61467) / 0.96343) + \varepsilon_i$$

On obtient l'ajustement suivant :



Ajustement de la courbe avec R , puis avec Sigmaplot.

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 F (les résultats sont donnés en annexe). 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.

Outre le phénomène d'hétéroscédasticité des résidus observé sur bon nombre d'échantillons, le diagnostique des régressions révèle un autre problème : celui des valeurs trop élevées des facteurs

d'inflation de la variance(VIF).Les VIF ont été obtenus avec sigmaplot. Pour l'échantillon **21A_PR3,** on obtient le tableau suivant :

Paramètres	VIF
c	41092.1918
d	411580.4782
e	503355.6899
g	72733.4919
a	39.0474
b	39.2517
phi	51.2958
tau	50.4516

Il a été déterminé de manière empirique pour un modèle à beaucoup de paramètres qu'un VIF supérieur à 10 était révélateur de multicolinéarité entre ces derniers.

Le VIF indique l'inflation de la variance des coefficients en présence de multicolinéarité.

Un facteur d'inflation de la variance trop élevé révèle une régression « instable ». Le moindre changement au niveau des paramètres (suppression ou ajout) modifie de manière excessive les estimations précédemment effectuées. De plus, certains paramètres peuvent être estimés avec un signe opposé à celui auquel on s'attendait. En résumé, les VIFs trop élevés sont révélateurs d'une régression à laquelle on ne peut pas se fier. Il convient donc d'éliminer des paramètres.

Choisir un nombre élevé de paramètre pour décrire le modèle permet de l'approcher de manière plus fine, au détriment de la précision de l'estimation de ceux-ci.

Il faut se limiter à des paramètres correspondant aux aspects essentiels du phénomène pour l'étude en cours.

Il va donc falloir éliminer les paramètres responsables du phénomène de multicolinéaité.

c) Matrice de corrélations entre les paramètres

Dans un premier temps, on cherche à les identifier grâce à la matrice de corrélation qu'on obtient en utilisant la fonction **« gnls »** de R, présente dans la librairie **« nlme »**, obtenue grâce à la fonction **« library »** de R.

Toujours dans le cas du même échantillon, on obtient la matrice de corrélation suivante :

	c	d	e	g	a	b	phi
d	-0.998						
e	0.994	0.999					
g	-0.987	0.994	-0.999				
a	0.569	-0.567	0.565	-0.562			
b	0.569	-0.569	0.568	-0.568	0.987		
phi	-0.089	0.065	-0.042	0.021	-0.034	-0.029	
tau	-0.049	0.025	-0.004	-0.017	-0.013	-0.008	0.989

Les matrices de corrélation sont du même type pour toutes les régressions.

Rappelons l'équation sur laquelle on travaille :

Y_i =c+(d*x_i)+(e*(x_i^2))+(g*(x_i^3))+(a-b*x_i)*sin(2*pi*(x_i+phi)/tau)+ε_i

Dans cette équation, les paramètres c, d e et g sont fortement corrélés entre eux. C'est également le cas des paramètres a et b et des paramètres phi et tau.

On a identifié les paramètres responsables de la colinéarité. On peut désormais simplifier l'équation en éliminant les paramètres superflus afin de réduire le phénomène de multicolinéarité. Cela se traduira par une diminution des VIF.

III) Elimination des paramètres responsables de la multicolinéarité.

L'objectif de cette étape était d'obtenir une équation simplifiée qui conviendrait à toutes les régressions, c'est-à-dire qui permettrait d'obtenir un bon ajustement, donc de bons indicateurs de régression (un F et un R^2 élevés) tout en diminuant la colinéarité entre les paramètres et donc en diminuant la valeur des VIF . Choisir un nombre élevé de paramètres pour décrire le modèle permet de l'approcher de manière plus fine, au détriment de la précision de l'estimation des paramètres. Il convient de se limiter aux paramètres correspondant aux aspects essentiels de l'étude en cours.

On a abordé ce problème de trois manières différentes. Dans un premier temps, on a supprimé des paramètres de manière aléatoire par le procédé suivant :

On a supprimé le paramètre qui était le plus responsable de la colinéarité, puis dans la nouvelle équation, on a recommencé le même procédé, jusqu'à l'obtention d'une équation optimale.

On a obtenu l'équation suivante :

Y_i =c+(e*(x_i^2))+(a-b*x_i)*sin(2*pi*(x_i+phi)/tau)+ε_i

Cette équation satisfait aux conditions précédemment énoncées pour quasiment toutes les régressions.

Cependant, il est nécessaire d'appliquer une méthode plus rigoureuse pour corroborer l'hypothèse faite pour la nouvelle équation.

On a appliqué à chaque série de données d'échantillon la méthode de la régression descendante. Celle-ci consiste à calculer les F de Fisher de chacun des paramètres et à éliminer celui dont le F de Fisher est le moins significatif. On reproduit ce procédé jusqu'à ce que tous les paramètres aient un F de Fisher significatif. Comme on a constaté que le paramètre phi était souvent non significatif, on a choisi de le remplacer dans chaque équation par la valeur à laquelle il est estimé.

Comment calcule-t-on le F de Fisher pour un paramètre ?

Dans le cas où on veut comparer un modèle restreint à un modèle complet, on utilise la statistique de test suivante :

$$F=\frac{(RRSS-URSS)/r}{URSS/(n-p-1)}$$

-Un modèle sans r variable est appelé un modèle restreint

- RRSS (Restricted Residuals Sum of Square)= somme des carrés des résidus du modèle restreint

-URSS (Unrestricted Residual Sum of Square)=somme des carrés des résidus du modèle complet

-MSE(Mean Square Error)= <u>somme des carrés des erreurs du modèle contenant tous les paramètres.</u>

Dans le cas d'une seule variable, r vaut 1, et donc en passant aux sommes de carrés du modèle on obtient :

$$F=\frac{(RRSS-URSS)/1}{URSS/(n-p-1)}=\frac{SS_{modèle\ complet}-SS_{modèle\ sans\ j}}{MSE}$$

Prenons à nouveau l'exemple de l'échantillon **21A_PR3**. Appliquons-lui la régression descendante.

On calcule le F de Fisher pour chacun des paramètres en utilisant la formule ci-dessus.

On commence afficher la régression complète puis on relève la valeur de la somme des carrés du modèle, puis on enlève le paramètre c et on relève la valeur de sa somme des carrés du modèle auquel on a ôté c.

Dans Sigmaplot, on obtient une analyse de la variance pour le modèle complet :

Corrected for the mean of the observations:

	DF	SS	MS	F	P
Regression	6	2480.4278	413.4046	661.9654	<0.0001
Residual	281	175.4876	0.6245		
Total	287	2655.9154	9.2541		

On obtient une analyse de la variance pour le modèle restreint :

	DF	SS	MS	F	P
Regression	5	2472.6076	494.5215	760.7697	<0.0001
Residual	282	183.3079	0.6500		
Total	287	2655.9154	9.2541		

$SS_{modèle\ complet}$= 2480.4275

$SS_{modèle\ sans\ c}$= 2472.6076

MSE=0.6245

On calcule $F_c=\frac{2480.4275-2472.6076}{0.6245}=\frac{7.8199}{0.6245}$=12.52186

Donc F_c=12.52186

Or $F_{1,282}$=3.87<12.52186

c a donc un apport significatif. Il ne sera pas éliminé.

On calcule les F de Fisher des autres paramètres de la même façon et on les reporte dans le tableau suivant :

	estimation	$SS_{\text{modèle complet}}$	SS_j	MSE	F
C	-33.2238	2480.4275	2472.6076	0.6245	12.52186
D	65.6056	2480.4275	2417.8261	0.6245	100.2424
E	-15.0378	2480.4275	2417.6736	0.6245	100.4866
G	1.0716	2480.4275	2423.6536	0.6245	90.91097
A	3.0290	2480.4275	2445.4619	0.6245	55.98975
B	0.4034	2480.4275	2467.8568	0.6245	20.12922
Tau	0.9630	2480.4275	2303.3463	0.6245	283.5568

Dans cet exemple, tous les paramètres ont un apport significatif. Ceci était prévisible aux vues des valeurs des statistiques t de Student dans la régression. Tous les paramètres étaient significativement différents de zéro.

S'il y avait eu un paramètre non significatif, on l'aurait supprimé et on aurait recommencé le même procédé sur la nouvelle équation obtenue, et ceci jusqu'à l'obtention d'un modèle dont tous les paramètres sont significatifs.

Tau a toujours un apport significatif, aussi ne l'a-t-on ôté pour aucune régression. Ceci nous arrange car l'étude porte sur la valeur de l'estimation du paramètre tau.

Après avoir éliminé les paramètres superflus de l'équation pour les 120 régressions à l'aide d'une régression descendante, on a fait de même à l'aide d'une régression ascendante et on a constaté qu'on n'obtenait pas toujours la même équation. On a donc gardé celle qui garantissait les valeurs de F et de R^2 les plus élevées et les valeurs de VIF les plus basses.

Pour la méthode ascendante, on introduit les paramètres un à un. On commence par un modèle à un paramètre et on introduit les autres paramètres un à un. On introduit à chaque pas le paramètre susceptible d'augmenter le plus la somme des carrés du modèle. Donc, le paramètre introduit en premier est celui qui possède la plus grande valeur de F et qui est significatif avec une probabilité par défaut associée au F de 0.5. Ce seuil s'appelle le seuil « pour entrer ».

On a constaté que malgré l'élimination d'un certain nombre de paramètres, les valeurs des VIF demeuraient trop élevées pour la plupart d'entre eux. Malgré tout, il faut faire attention à ne pas trop en supprimer. En effet, bien que la suppression de ceux-ci induise la diminution de la valeur des VIF, elle provoque également une perte d'information d'un point de vue biologique. Il faut donc concilier les exigences des domaines mathématiques et biologiques afin d'obtenir des résultats optimaux. Enfin, il faut tenir compte du fait que ces méthodes de régressions ne donnent pas nécessairement le meilleur modèle. Elle ne sont donc pas entièrement fiables.

Eliminer des paramètres a permis d'améliorer la statistique du test de Bartlett pour certaines régressions. Les anciens résultats ainsi que les nouveaux sont donnés en annexe.

IV) Calcul des nouvelles périodes

On a désormais éliminé tous les paramètres superflus. Pour chaque échantillon, on a établi une nouvelle équation. On a donc de nouvelles estimations du paramètre « **Tau** ». Il faut alors recalculer les valeurs de la période dans chacun des 120 échantillons.

On a donc pris en compte les nouvelles estimations de « **Tau** », et on les a multipliées par 24 pour que la période soit exprimée en jours.

Prenons l'exemple de l'échantillon 19A.INL2. On a obtenu une nouvelle équation par la méthode de la régression descendante. Avant, on obtenait une estimation pour les 8 paramètres pour l'équation

Y_i =c+(e*(x_i^2))+(a-b*x_i)*sin(2*pi*(x_i+phi)/tau)+ε_i

paramètres	estimation	VIF
C	192.9826	2553.6215
D	6.5844	28725.5335
E	-17.6857	42541.3837
G	2.6456	7469.3863
A	41.7862	15.8591
B	8.7904	15.5278
Phi	0.3260	24.8913
Tau	1.0300	23.3141

On obtenait donc une estimation de « **tau** » égale à 24.72 heures. On constate également qu'on obtient des VIF excessifs.

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

y=c+e*x^2+g*x^3+(a-b*x)*sin((2*pi*(x+0.32))/tau)

ainsi que de nouvelles estimations pour les paramètres c, e, g, a, b et tau :

paramètres	estimations	VIF
C	198.8907	24.7934
E	-15.3813	346.3458
G	2.3898	218.5831
A	42.0590	13.2943
B	8.8712	13.5270
Tau	1.0296	1.0365

On a obtenu une estimation de « **tau** » égale à 24.7104 heures, qui est assez proche de l'estimation précédente. La suppression de paramètres a permis de diminuer les valeurs des VIF, bien que celles-ci restent élevées du fait des liaisons qui existent entre c, e et g, et entre a et b.

Avant les suppressions de paramètres, on obtenait les huit périodes moyennes suivantes :

Sample	Period	Mean std error	Stand dev
Whole Retina	22.941	0.035	0.416
GCL+INL	23.425	0.035	0.295
GCL	26.352	0.286	2.332
INL	26.378	0.261	2.066
PRL	26.316	0.488	1.838
PRL+INL	28.265	0.846	3.563
Dissociated Cells	29.095	0.256	1.250
CoGC+INL	27.046	0.084	0.716

Désormais, on obtient celles-ci :

Sample	Period	Mean std error	Stand dev
Whole Retina	22.888	0.018	0.457
GCL+INL	23.418	0.010	0.296
GCL	25.974	0.044	1.747
INL	26.254	0.059	2.089
PRL	26.318	0.280	1.850
PRL+INL	27.462	0.225	2.690
Dissociated Cells	29.058	0.256	1.250
CoGC+INL	27.036	0.024	0.713

Globalement, on obtient des périodes moyennes assez proches de celles calculées au départ, sauf pour les groupes **GCL** et **PRL+INL**. On trouvait en effet initialement dans ces groupes des valeurs de périodes très élevées (voir annexe) .Par exemple, dans le cas de l'échantillon **27E.PR+INL8** du groupe **PRL+INL ,** on avait une période égale à 32,96 heures. On a désormais une période égale à 27.4896, qui est une valeur bien plus proche de celles des autres échantillons du même groupe. Dans le groupe **GCL,** la période de l'échantillon **10B.GC5** est passée de 32.1912 heures à 26.2848 heures.

On constate également que la moyenne de l'erreur type est diminuée pour tous les groupes(voir annexe).

La suppression des paramètres superflus aura donc permis d'améliorer la qualité des régressions, ainsi que d'augmenter dans certains cas la valeur de la statistique de Bartlett. Elle aura également permis de diminuer les valeurs de périodes qui étaient trop élevées par rapport à celles du reste du groupe, ainsi que de diminuer la valeur de la moyennes de la somme des carrés des erreurs.

V)Recherche d'une période commune au sein de chaque groupe

L'objectif de cette partie est de mettre en évidence l'existence d'une période homogène entre les différents échantillons d'un même groupe, c'est-à-dire que les périodes observées sont globalement assez proche pour qu'on considère que le groupe est cohérent. Pour vérifier cela, on veut montrer qu'on peut, pour chaque groupe, ajuster une période T_0 unique.

1°)Présentation de la méthode

Soit n le nombre d'échantillons contenus dans le groupe.

On souhaite comparer deux modèles d'ajustement :

Pour le premier modèle , on ajuste la fonction y=f(c_i, d_i , e_i , g_i , a_i , b_i , phi_i , T_0) , où T_0 $est\ fixée$ avec $1 \leq i \leq n$ à toutes les données.

Avec :

$$\begin{cases} c = c_i \\ d = d_i \\ e = e_i \\ g = g_i \\ a = a_i \\ b = b_i \\ phi = phi_i \end{cases}$$

Pour le second modèle, on ajuste la fonction y=f(c_i, d_i , e_i , g_i , a_i , b_i , phi_i , T_i) avec $1 \leq i \leq n$ à toutes les données

Avec

$$\begin{cases} c = c_i \\ d = d_i \\ e = e_i \\ g = g_i \\ a = a_i \\ b = b_i \\ phi = phi_i \\ T = T_i \end{cases}$$

On est dans une situation où l'on doit comparer deux modèles : un modèle complet et un modèle restreint.

On utilise la statistique de test suivante :

$$F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1)}$$

Avec :

-RRSS (Restricted Residual Sum Of Square) = Somme des carrés des résidus du modèle restreint

-URSS(Unrestricted Residual Sum Of Square) = Somme des carrés des résidus du modèle complet

r est le nombre de paramètres que l'on a ôté au modèle complet.

Le modèle complet contient p paramètres

F suit une loi de Fisher à r,n-p-1 degrés de liberté.

On va comparer ces modèles pour chacun des huit groupes en calculant la statistique de Fisher qui leur est associée.

Etant donné la taille de certains échantillons, nous donnerons les résultats détaillés pour deux d'entre eux. Les résultats finaux pour les six autres groupes seront donnés à la fin de cette partie.

Le code R est donné en annexe.

2°)Recherche d'une période commune pour les échantillons du groupe CoGC+INL

On effectue deux régressions communes pour les 7 échantillons du groupe. On effectue tout d'abord une régression restreinte, puis on effectue une régression complète. Ensuite, pour chaque modèle, on calcule les valeurs de URSS, RRSS, r et n-p-1. Enfin, on calcule la valeur de la statistique F de Fisher.

On travaille sur l'équation suivante : $y = c + d \times x + a \times \sin\left(\frac{2 \times \pi \times x}{tau}\right)$

Dans R, on obtient

Paramètres	Estimation	Std Error	t value	Pr(>|t|)
c1	126.538854	6.01251	21.064	$< 2 \times 10^{-16}$
c2	127.059637	6.037339	21.046	$< 2 \times 10^{-16}$
c3	127.525492	6.050340	21.077	$< 2 \times 10^{-16}$
c4	128.524011	6.050943	21.240	$< 2 \times 10^{-16}$
c5	128.355672	6.065301	21.162	$< 2 \times 10^{-16}$
c6	127.563023	6.050974	21.081	$< 2 \times 10^{-16}$
c7	127.634761	6.051560	21.091	$< 2 \times 10^{-16}$
d1	-10.538951	1.717851	-6.135	1.02×10^{-09}
d2	-10.686915	1.727999	-6.185	7.53×10^{-10}
d3	-10.814142	1.726271	-6.264	4.57×10^{-10}
d4	-11.115816	1.726606	-6.438	1.51×10^{-10}
d5	-11.038528	1.725278	-6.398	1.95×10^{-10}
d6	-10.832858	1.726580	-6.274	4.30×10^{-10}
d7	-10.838817	1.726528	-6.278	4.20×10^{-10}
a1	7.196868	2.155076	3.339	0.000855
a2	7.171047	2.163271	3.317	0.000927
a3	7.116742	2.165054	3.290	0.001020
a4	6.886031	2.166202	3.181	0.001493
a5	6.930913	2.161862	3.200	0.001398
a5	7.500471	2.161671	3.469	0.000233
a7	7.405012	2.162014	3.426	0.000626
Tau	1.061834	0.005679	186.971	$< 2 \times 10^{-16}$

Tableau de la régression commune restreinte pour les sept échantillons du groupe **CoGC+INL**

paramètres	estimations	Std Error	t value	Pr(> |t|)
c1	126.64082	6.12422	20.679	2×10^{-16}
c2	127.04509	6.14828	20.664	2×10^{-16}
c2	127.54475	6.16682	20.682	2×10^{-16}
c4	128.46897	6.17121	20.817	2×10^{-16}
c5	128.30420	6.19080	20.725	2×10^{-16}
c6	127.60359	6.16430	20.700	2×10^{-16}
c7	127.59329	6.16748	20.688	2×10^{-16}
d1	-10.57044	1.75461	-6.024	2.02×10^{-09}
d2	-10.68244	1.76423	-6.055	1.67×10^{-09}
d3	-10.82008	1.76458	-6.132	1.05×10^{-09}
d4	-11.09883	1.76629	-6.284	4.05×10^{-10}
d5	-11.02260	1.76696	-6.238	5.39×10^{-10}
d6	-10.84537	1.76458	-6.149	9.42×10^{-10}
d7	-10.82605	1.76384	-6.135	1.02×10^{-09}
a1	7.20003	2.15858	3.336	0.000867
a2	7.17102	2.16525	3.312	0.000943
a3	7.11702	2.16660	3.285	0.001038
a4	6.88604	2.16817	3.176	0.001516
a5	6.93061	2.16930	3.195	0.001421
a6	7.50118	2.16511	3.464	0.000543
a7	7.40506	2.16486	3.421	0.000638
Tau1	1.06310	0.01502	70.766	2×10^{-16}
Tau2	1.06165	0.01508	70.396	2×10^{-16}
Tau3	1.06207	0.01518	69.980	2×10^{-16}
Tau4	1.06114	0.01565	67.795	2×10^{-16}
Tau4	1.06121	0.01553	68.330	2×10^{-16}
Tau6	1.06231	0.01441	73.735	2×10^{-16}
Tau7	1.06135	0.01459	72.759	2×10^{-16}

Tableau de la régression commune complète pour les sept échantillons du groupe **CoGC+INL**

Tableaux d'analyse de la variance pour les sept échantillons du groupe CoGC+INL

Effets	Degrés de liberté	Somme des carrés	Moyenne de la somme de carrés
Régression	21	215436	10258.8
Résidus	1996	1281103	641.8
Total	2017	1496539	741.98

Tableau d'analyse de la variance pour le modèle restreint

Effets	Degrés de liberté	Somme des carrés	Moyenne de la somme des carrés
Régression	27	215444	7979.40
Résidus	1990	1281095	643.76
Total	2017	1496539	741.98

Tableau d'analyse de la variance pour le modèle complet

RRSS=215444

URSS=215436

Dans le modèle complet, on estime 28 paramètres, tandis qu'on en estime 22 dans le modèle restreint.

r=28-22=6

On a 2017 observations, et on estime 28 paramètres dans le modèle complet.

n-p-1=2017-28-1=1988

Composantes de la statistique de test	**Valeur des composantes**	**Statistique de test**
RRSS	1281103	0.00206
URSS	1281095	
r	6	
n-p-1	1988	

On obtient une valeur de F de Fisher égale à 0.00206. Or, $F_{6,1988}$=2.80>0.00206. On ne rejette donc pas l'hypothèse H_0 : Il n'existe aucune différence entre les deux ajustements pour chacune de nos données au sein d'un même groupe.

Il est donc possible d'ajuster une période unique pour le groupe **CoGC+INL.**

Voici les résultats des tests de Fisher pour les autres groupes :

Groupe	Résultat du test de Fisher	Valeur relevée dans la table de valeurs de la statistique de Fisher	Conserve-t-on H_0?
PRL	0.1185	1.83	Oui
GCL	0.0018	1.67	Oui
INL	0.28	2	Oui
GCL+INL	0.008	1.79	Oui
PRL+INL	0.0076	1.72	Oui
Dissociated cells	0.011	1.71	Oui
Whole retina	0.0067	1.60	Oui

On conserve l'hypothèse H_0 dans chacun des huit tests. On peut donc ajuster une période unique T_0 pour chaque échantillon au sein de chaque groupe. Il existe donc une cohérence du point de vue de la période au sein de chaque groupe.

On sait désormais que les groupes sont homogènes au niveau de la période. Il s'agit à présent de les comparer afin de pouvoir conclure quant à la situation de l'horloge biologique de la rétine.

VI) Comparaison entre les groupes

Afin de mieux comprendre le fonctionnement de l'horloge biologique au sein de la rétine, on va effectuer les comparaisons entre les périodes des groupes suivants :

Groupes	Rétine entière	PRL	INL	GCL	PRL+INL	GCL+INL	Cellules dissociées	CoGC+ INL
Rétine entière								
PRL	A comparer							
INL	A comparer	A comparer						
GCL	A comparer	A comparer	A comparer					
PRL+INL	A comparer	A comparer	A comparer					
GCL+INL	A comparer		A comparer	A comparer				
Cellules dissociées		A comparer	A comparer	A comparer				
CoGC+INL			A comparer	A comparer	A comparer			

Les cellules rouges indiquent des comparaisons qui ont déjà été mentionnées ailleurs dans le tableau

Les cellules noires indiquent des comparaisons qu'il est inutile de réaliser.

Il y a 18 comparaisons à effectuer. On peut réduire ce nombre si les périodes des groupes GCL, PRL et INL ne sont pas significativement différentes. On pourra former un seul groupe à partir de ces trois groupes et le comparer à d'autres groupes, plutôt que de comparer chacun de ces trois groupes individuellement.

On aimerait réaliser ces comparaisons par le biais d'ANOVAs, mais il s'avère que chaque période est obtenue avec une erreur type (voir annexe). Il faudrait donc montrer que les erreurs sont négligeables pour chaque échantillon, afin de pouvoir considérer que la valeur de période obtenue est la valeur exacte.

L'objectif de cette partie est donc :

-De montrer qu'on aboutit à la même conclusion par une régression générale et par une ANOVA

-De montrer que la variance intra-groupe est négligeable par rapport à la variance inter-groupe

Dans un premier temps, on effectue ces comparaisons en faisant une régression générale sur les données des différents groupes à comparer, d'abord en ajustant une valeur de Tau unique, puis en ajustant n valeurs de Tau et en comparant les deux modèle comme précédemment. On calcule ensuite la statistique F de Fisher de la même manière que dans la partie V). On teste :

H_0 : Il y a une homogénéité de la période au sein du groupe

Contre

H_1 : Les périodes au sein du groupe ne sont pas homogènes

On a déjà démontré que les périodes étaient homogènes au sein d'un même groupe. Ainsi, si à l'issue du test on conclut à la conservation de l'hypothèse H_1 , on saura qu'il y a une différence significative entre les périodes des deux groupes comparés.

On a commencé par comparer les résultats obtenus avec la régression générale avec ceux obtenus par l'ANOVA sur les périodes des groupes faite en négligeant les erreurs type.

On a comparé les groupes **GCL** et **INL**

On obtient les résultats suivants :

1°) Avec une régression générale

On regroupe les données des échantillons **GCL** et **INL** et on teste l'hypothèse H_0 citée ci-dessus contre l'hypothèse H_1 citée ci-dessus .

Tableaux d'analyse de la variance pour les 33 échantillons des groupes GCL et INL

Effets	Degrés de liberté	Somme des carrés	Moyenne de la somme des carrés
Régression	66	649200	9836.36
Résidus	9534	25847368	2711.07
Total	9600	26496568	2760.049

Tableau d'analyse de la variance pour le modèle restreint

Effets	Degrés de liberté	Somme des carrés	Moyenne de la somme des carrés
Régression	98	654403	6677.58
Résidus	9502	25842165	2719.65
Total	9600	26496568	2760.049

Tableau d'analyse de la variance pour le modèle complet

RRSS= 654403

URSS=649200

r=32

n-p-1=9600-99-1=9500

F=0.017

$F_{32,9500}$ =1.46

Composantes de la statistique de test	Valeur des composantes	Statistique de test
RRSS	25847368	**0.059**
URSS	25842165	
r	32	
n-p-1	9500	

On obtient une valeur de F de Fisher égale à 0.059$< F_{32,9500} = 1.46$. On ne rejette donc pas H_0 : il n'y a pas de différence significative au niveau des périodes au sein du groupe. On ne peut donc pas conclure à une différence significative de période entre **GCL** et **INL**.

2°)Avec la fonction « aov » de R

On teste les hypothèses suivantes :

H_0 : Les périodes des deux groupes ne sont pas significativement différentes

Contre

H_1 : Il y a une différence significative entre les périodes des deux groupes.

La statistique de test utilisée est la suivante :

$$F=\frac{\frac{Somme\ des\ carrés\ moyens\ dus\ au\ facteur}{degré\ de\ liberté\ associé}}{\frac{somme\ des\ carrés\ résiduels}{degré\ de\ liberté\ associé}}$$

On vérifie tout d'abord qu'on a bien l'homoscédasticité et la normalité des résidus par des tests de Bartlett et de Shapiro-Wilk.

```
Bartlett test of homogeneity of variances

Data: res by groups

Bartlett's  K-squared=0.4745 , df=1 , p-value=0.4909
```

On ne rejette pas l'hypothèse H_0 d'homogénéité des variances des résidus.

Shapiro-Wilk normality test

Data : res

W=0.9661 , p-value= 0.3801

On ne rejette pas l'hypothèse H_0 de normalité des résidus.

Effets	**Degrés de Liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	**Pr(> F)**
Groupe	1	0.646	0.646	**0.1733**	**0.6801**
Résidus	31	115.568	3.728		

Tableau d'analyse de la variance pour la comparaison entre les périodes du groupe GCL et du groupe INL à partir des données des tableaux en annexe

On aboutit au même résultat que par la méthode de la régression générale, on conserve l'hypothèse H_0 : il n'y a pas de différence significative entre les deux groupes.

Pour la comparaison des périodes des groupes **GCL** et **INL**, on aboutit à la même conclusion par les deux méthodes.

3°) Présentation d'une méthode permettant de négliger les erreurs associées à la période

On a ensuite cherché à montrer que les variance intra-échantillon étaient négligeables face aux variance inter-échantillons au sein de chaque groupe.

Rappelons le résultat fondamental de l'ANOVA :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y})^2 = J\sum_{i=1}^{I}(\bar{y_i}-\bar{y})^2 + \sum_{i=1}^{I}(\sum_{j=1}^{J}(y_{ij}-\bar{y_i})^2)$$

Qui s'écrit encore :

Somme des carrés totaux=Somme des carrés dus au facteur+somme des carrés dus aux résidus

Ou

Variation totale=variation inter-groupe+variation intra-groupe

On va considérer les quantités « variation inter-groupe » et « variation intra-groupe » et calculer leur rapport pour chacun des huit groupes.

Les résultats sont donnés sous forme de tableaux d'analyse de la variance.

GROUPE PR					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{23,3563}$
Inter-échantillon	23	37966	1650.69	**11.8**	**1.54**
Intra-échantillon	3563	497831.80	139.72		
Totale	3586	535603.90	149.35		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe **PR**

GROUPE GC					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{65,4539}$
Inter-échantillon	65	254755	3919.307692	**2.83**	**1.31**
Intra-échantillon	4539	6266690	1380.63		
Totale	4604	6521445	1416.47		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe **GC**

GROUPE INL					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{52,4494}$
Inter-échantillon	52	482441	9277.71	**2.44**	**1.35**
Intra-échantillon	4494	17039736	3791.66		
Totale	4996	17522177	3507.24		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe **INL**

GROUPE DISSOCIATED CELLS					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{29,4002}$
Inter-échantillon	29	68257	2353.68	**11.5**	**1.45**
Intra-échantillon	4002	817194	204.196		
Totale	4031	885451	219.66		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe des cellules dissociées

GROUPE PRL+INL					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne des la somme des carrés**	**F**	$F_{57,3982}$
Inter-échantillon	57	360166	6318.70	**5.656**	**1.32**
Intra-échantillon	3982	4447996	1117.02		
Totale	4039	4808162	1190.43		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe des **cellules dissociées.**

GROUPE GCL+INL					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{45,3125}$
Inter-groupe	45	8922166	198270.35	**8.75**	**1.37**
Intra-groupe	3125	70756566	22642.10		
Totale	3169	79678732	25143.17		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe **GCL+INL**

GROUPE CoGC+INL					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{28,1989}$
Inter-groupe	28	1496539	53447.8	**83.006**	**1.48**
Intra-groupe	1989	1280833	643.9		
Totale	2017	1496539	741.96		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe **CoGC+INL**

GROUPE WHOLE RETINA					
Variation	**Degrés de liberté**	**Somme des carrés**	**Moyenne de la somme des carrés**	**F**	$F_{110,6238}$
Inter-groupe	110	342444294	3113129.94	**11.09**	**1.24**
Intra-groupe	6238	1750461191	280612.56		
Totale	6348	2092905485	329695.25		

La variance intra-échantillon est négligeable par rapport à la variance inter-échantillon dans le groupe des **rétines entières**

On a montré qu'on pouvait, dans chaque groupe, négliger la variance intra-groupe.

On va donc effectuer des comparaisons entre les périodes des groupes qui nous intéressent en considérant que les périodes sont des données exactes, pour lesquelles il n'y a pas d'erreur standard associée. On effectue ces comparaisons par le biais d'ANOVA avec la fonction « AOV » de R.

On donne les p-value de ces ANOVAS pour chaque comparaison. On a à chaque fois effectué des tests de Bartlett et de Shapiro-Wilk sur les résidus qui nous ont conduits à ne pas rejeter les hypothèses H_0 d'homogénéité et de normalité.

Rappelons les hypothèses testées :

H_0 **: Les périodes des deux groupes ne sont pas significativement différentes**

Contre

H_1 **: il existe une différence significative de période entre les deux groupes**

4°) Résultats des ANOVA

groupes	Rétine entière	PRL	INL	GCL	PRL+INL	GCL+INL	Cellules dissociées	CoGC+INL
Rétine entière								
PRL	2.238×10^{-9} H_1							
INL	9.377×10^{-9} H_1	0.6801 H_0						
GCL	1.897×10^{-9} H_1	0.6286 H_0	0.8758 H_0					
PRL+INL	3.904×10^{-4} H_1	0.1128 H_1	0.1128 H_0					
GCL+INL	0.001494 H_0		3.332×10^{-5} H_1	3.332×10^{-5} H_1				
Cellules dissociées		0.0002030 H_1	0.0001330 H_1	7.384×10^{-6} H_1				
CoGC+INL			0.3492 H_0	0.1394 H_0	0.6888 H_0			

Il n'y a pas de différence significative entre les périodes des groupes GCL PRL et INL

Il y a une différence significative entre les périodes du groupe GCL PRL INL et celle de la rétine entière

Il y a une différence significative entre les périodes du groupe GCL PRL INL et celle des cellules dissociées

Il y a une différence significative entre les périodes de la rétine entière et du groupe GCL+INL

Il y a une différence significative entre les périodes du groupe GCL+INL et celle du groupe GCL INL

Il y a une différence significative entre les périodes de la rétine entière et celle du groupe PRL+INL

Il n'y a pas de différence significative entre les périodes du groupes PRL+INL et celle du groupes PRL INL.

VII) Conclusion

Le but de l'étude était de déterminer où se situait l'horloge biologique de la rétine.

Dans un premier temps, on a isolé les différentes couches de la rétine afin d'étudier leur comportement.

Les couches **GCL INL** et **PRL** ont toutes les trois une période similaire, cependant leur période est différente de celle de la **rétine entière**. Elle est plus importante.

Les couches gardent donc une activité, même lorsqu'elles sont isolées les unes des autres, cependant leur période est différente de celles qu'elles ont lorsqu'elles forment la rétine entière.

Ensuite, on a recollée les couches de la rétine afin de voir si la nature de leur communication avait changé. On a également crée les groupes **GCL+INL** et **PRL+INL** en ôtant respectivement les couches **PRL** et **GCL** de la rétine.

Les couches **GCL PRL** et **INL** séparées ont une période différente de celle des cellules dissociées.

La rétine entière a une période différente de celle du groupe où l'on a ôté la couche **PRL** et de celle du groupe où l'on a ôté la couche **GCL**

Le groupe **GCL+INL** a une période différente de celle du groupe **GCL INL**

En revanche, le groupe **PRL+INL** a une période comparable à celle du groupe **PRL INL**. On n'a pas encore trouvé d'explication à ceci.

La période est modulée par le degré de couplage entre les oscillateurs et les cellules.

Ces analyses permettront à C.Jaeger de tirer des conclusions plus fines sur la situation de l'horloge dans la rétine et sur la nature de la communication entre les couches.