Validation des méthodes d'identification bactérienne.

Présenté par: ABOU SENE

Master2 S.T.A.F.A.V UFR S.A.T

Université Gaston Berger
STAGE effectué au:
Laboratoire bacteriologie-virologie fondamentale et appliquée
de l'U.C.A.D II

20 octobre 2008

Table des matières

Résumé

La prédiction des aptitudes de croissance des micro-organismes en vue d'une prise de décision est un problème couramment rencontré dans le domaine médical. Cette décision, le plus souvent associée à un taux de risque est en général soumise à des contraintes économique et temporelles. Dans le contexte actuel où les technologies et la recherche sont de plus en plus affinées, des méthodes de prédiction permettant avec un moindre coût une identification rapide et efficace sont de rigueur.

La première partie de ce travvail traite de la généralité dans le cadre de l'étude. Après avoir présenter les données, on retrace les différentes méthodes utilisées pour leurs acquisitions tout en mettant en évidence les erreurs de mesures associées.

La deuxième partie met l'accent sur la prédiction d'une cinétique de croissance bactérienne à partir de ses premiers points expérimentaux. Les méthodes de la modélisation classique ne permettant toujours pas de répondre à cette problématique, d'autres altérnatives comme les techniques d'inférence bayésienne peuvent servir à des éléments de réponses.

Dans la troisième partie on essaie de développer une approche modélisatrice de la concordance des caractères biochimiques en fonction du temps d'incubation et de la taille des inocula.

Mots clés : temps d'incubation, inoculum, prédiction, concordance

Abstract

In medical microbiology, one often needs to predict growth abilities of microorganisms in order to take a decision. This decision is often associated with a serious risk and generally constrained by economic and time requirements. Actually context, where technology and research are more and more refined developping prediction methods quickly giving the identification at a low cost is the rigour.

The first part of this work deals with generally of the study context. After describing the data, we give different methods use for there procuration by making out errors measures associated.

The second part deals with the prediction of the growth kitnetics of a population of microorganisms from its experimental points. The classsical modelling methods do not alway permit to solve this problematical, others alternative like population approach based on bayesian inference can serve to elements of answers.

In the third part, one tries to develop and modelling approach of the concordance of the biochimical characters according to the time of incubation and the size of inoculated them.

Key words : time of incubation, inoculated, prediction, concordance

Introduction

La prédiction des aptitudes de croissance des micro-organismes soumis à des conditions données est un problème d'actualité qui touche beaucoup de secteurs notamment le secteur de la microbiologie médicale. Dans ce dernier on s'intéresse particulièrement aux micro-organismes patohgénes, responsables de maladies infféctueuses. Dans le monde médical comme dans les autres secteurs de la microbiologie, la prise de décision associée à la prédiction est souvent accompagnée d'une prise de risque importante. Fâce à tel danger, la prédiction se doit d'être la plus fiable possible afin de minimiser le risque.

De nos jours la prise de décision est devenue plus facile et plus fiable dans les laboratoires par l'identification des germes avec les galeries d'identifications, et le plus souvent par les galeries **API**. Le coût élevé de ces galeries importées pour les pays en voie de développement et dont le fondemment théorique reste encore un secret industriel a amené l'expertise locale à initier une serie de recherches dans les laboratoires univérsitaires l'**UCAD** pour aider les populations de moyens modèstes à accéder à des soins de qualité. C'est ainsi que le laboratoire bacteriologie-virologie fondamentale et appliquée, sous la direction du professeur *C.S. Boye* a mis en place des mini-galeries d'identification : les Microméthodes CSB. Il s'agit de plaques constituées de micro-cupules contenant des substrats déshydratés, ceux-ci permettant la mise en évidence d'activités enzymatiques ou d'assimilation de subtrats carbonés en milieu hostile ou approprié.
Même si l'efficacité et la fiabilté des Micro-CSB ont été prouvées par des études antérieures, ces dernières souffrent encore des problèmes de la standardisation de l'inoculum et du temps d'incubation des plaque de 24 H. Tout ceux-ci combinés, constituent un handicap dans l'établissement efficace et rapide d'un traitement. Dans un souci d'affiner et d'améliorer l'efficacité de ces micro-plaques, ce travail se donne comme objectif d'accélérer et de minimiser le coût de la prise de décision en développant des méthodes de prédiction d'une cinétique de croissance d'une population microbienne à partir de ses premiers points expérimentaux. Ainsi pour arriver à cet'objectif, par différentes méthodes, on va prédire les aptitudes de croissance d'une population bactérienne à partir de ses premiers points expérimentaux afin d'identifier l'inoculum adéquat permettant d'atteindre un taux de significativité maximale en un temps d'incubation aussi court que possible.

Chapitre 1

Généralités

1.1 Présentation de la structure d'accueil

La faculté de médecine, pharmacie d'odonto-stomatologie est régie par le décret numéro $70-1135$ du 13 octobre 1970 modifié. C'est un établissement public doté de personnalité juridique et d'autonomie financière, dirigé par un doyen élu. Elle a d'abord été l'école de médecine de Dakar crée en 1916, puis érigée en faculté en 1962 suite à la fondation officielle de l'université de Dakar le 24 février 1957. La faculté comprend trois sections : médecine, pharmacie, et odontologie (ex institut d'odonto-stomatologie). Elle est structurée en départements dirigés chacun par un chef de département élu.

Nous avons effectué notre stage à la section de pharmacie, au loboratoire bactériologie-virologie fondementale et appliquée placée sous l'autorité du doyen de la faculté. Il a pour buts (à suivre)

1.2 Données

L'étude porte sur les courbes de croissance de quatre espèces que sont :
Le Escherichia coli, le Klesbsialla pneumonia, Sallmonella paratyphi A et Acinetobacter. Leurs cultures en milieu non renouvelé (**batch**) avec des conditions de croissance idéales ont permi d'obtenir leurs croissances à des intervalles de temps reguliers par des techniques de dénombrement sur boites de petri.

1.2.1 Technique de mesure de la croissance

Parmi les techniques de mesures de la croissance, le dénombrement sur boîtes de petri constitue sans aucun doute la méthode la plus classique (*McMeekin et al.,* 1993). Cette technique permet la mesure de densités de populations comprises entre 10 et $10^8 - 10^9$ cellules/ml ou cellules/g de produit. Le principe repose sur l'hypothèse qu'une cellule viable déposée sur le gel nutritif de la boîte se divise jusqu'à l'obtention d'un amas de cellules issues de cette seule cellule-mére : une colonie.
En conséquence, le dénombrement des colonies revient au dénombrement des cellules ou groupes de cellules déposés et viables (raison pour la croissance est mesurée en

unités formant colonie : UFC). C'est une méthode très simple mais coûteuse en temps et en matériel et peu conduire à l'obtention de cinétiques constituées d'un nombre limité de points expérimentaux. Elle est identifiée comme étant la méthode la moins sensible aux erreurs de mesures, et elle ne tient compte que les cellules viables contrairement à la turbidimétrie, qui ne consiste pas à compter les cellules, mais à mesurer la biomasse microbienne, c'est à dire la masse de cellules séche totale.

1.2.2 Les erreurs de mesures

Les données sont obtenues par des expériences planifiées et controlées. Lorsque qu'on effectue une mesure, quelle qu'elle soit, l'instrument n'est pas toujours parfait. Le geste et la pratique ne sont toujours pas optimales. De ce fait le phénomène qu'on pense capter avec la mesure de x ne l'est qu'à un écart ou erreur prés $y = x + \varepsilon$.
De plus la mesure x qu'on effectue chez un individu au temps t, x_t varie au cours du temps comme dans tous les organismes vivants. Ainsi à la variabilité extrinsèque liée à la méthode utilisée, il y'a une variabilité intra-individuelle au cours du temps. En plus, entre différents sujets, il existe une variabilité inter-individuelle. Au total pour un mesure G d'une certaine grandeur de l'organisme, il existe plusieurs sources de variabilités :
1)Variabilité intra-individuelle.
2)Variabilité inter-individuelle.
3)Anomalie dans la mesure faite par un appariel de mesure.
4)Anomalie dans le geste de l'opérateur.

1.3 Situation de la problématique de prévision

Lorqu'un microbiologiste veut prédire l'évolution d'une population de micro-organismes contaminants, il réalise expérimentalement des cinétiques dans les conditions environnementales correspondant aux caractéristiques physico-chimiques de l'espèce. Si l'un des facteurs influençant la croissance comme la température ou le PH est modifié, le biologiste n'a plus qu'à refaire ses expériences dans les mêmes conditions considérées. Cette démarche est coûteuse et ne permet pas d'avoir une réponse rapide. Des microbiologistes on montré que l'on peut se servir d'expériences antérieures pour prédire le comportement de la flore microbienne en étudiant l'influence des principaux facteurs environnementaux sur la cinétique de croissance des micro-organismes (*McMeekin et Olley*, 1986 ; *Baird Parker et Kilsby* , 1987). Ainsi de nos jours est née une nouvelle discipline en pleine expansion dans le secteur de la microbiologie médicale commme dans le secteur de l'agro-alimentaire : **la microbiologie prédictive**. Elle consiste à développer des modèles reliant les paramètres de croissance des populations de micro-organismes aux principaux facteurs influençant la croissance. Ces modèles sont ensuite utilisés pour prédire l'évolution de la flore contamminante.

Principalement deux types de modèles sont souvent utilisés : le modèle prédictif probabiliste qui permet de prédire la probabilité avec laquelle l'événement entendu va se réaliser, et le modèle prédictif cinétique qui prédit le taux de croissance de

développement d'une population bactérienne.
Etant donné que c'est l'objectif de l'étude qui oriente le choix de modélisation, c'est ce dernier type de modèle que nous tenteront de développer dans notre étude.

1.4 L'outil de travail

L'outil priviliégé de travail est l'ordinateur. En effet, de nos jours on rencontre de plus en plus de grandes bases de données sur lesquelles on aimerait en tirer dans un bref délai le maximun d'information possible. De nombreux logiciels d'analyses statistiques sont maintenant à la disposition du staticien : **SAS**, **STATA**, **Splus**, **Ep-info**, **PS**, **SPSS**, **Eviews**, **R**,$\cdots$. Dans le cadre de ce travail, nous avons utilisé le logiciel **R** version 2.6.2. Depuis janvier 2008, la version 2.7.2 est disponible sur le site **CRAN** (Comprehensive **R** Archive Network). On a préféré utilisé la version 2.6.2 car l'essentiel des paquets dont on a besion y étaient déjà chargés.

R désigne à la fois un logiciel et un langage de programmation. Il s'agit d'un système qui permet de réaliser des analyses statistiques. C'est en particulier un outil bien adapté pour la manipulation des données, l'analye de données, l'utilisation de méthodes statistiques et les représentations graphiques. La conception de **R** a été largement influencée par **Splus** qui est un langage développé par les laboratoires *Bell* et commercialisé par la société *Insightful*. **R** est un logiciel libre, ce qui signifie que les utilisateurs ont la liberté d'exécuter, de copier, de distribuer et d'améliorer ce logiciel. **R** est disponible pour les systèmes d'exploitation *Unix, Linux, Windows, et Macosx* à l'adresse suivante : *http ://www.r-project.org/*

Chapitre 2

Prédiction de la cinétique de croissance d'une population de micro-organismes à partir de ses premiers points expérimentaux

2.1 Classification des modèles existant

Un modèle est une représentation simplifiée de la réalité, facilitant la prédiction ou l'estimation et est exprimé en langage mathématique. Toute la réalité ne pouvant pas être modélisée, un choix s'impose sur la partie de la réalité à modéliser. L'espace du modèle est alors défini, ceci implique que les informations qui pourront être extraites de la modélisation ne seront valables que pour l'espace choisi (*Tomassone et al.*,1993). Dans le domaine de la microbiologie, trois classes de modèles ont été proposés en 1993 par *Whiting et Buchnan* : les modèles primaires, secondaires et tertiaires.

Les modèles primaires (empiriques) décrivent l'évolution au cours du temps de la population microbienne dans un environnement spécifique. L'utlisation de ces modèles est essentiellement pratique : il s'agit par exemple de prédire le temps au bout du duquel la densité bactérienne aura dépassé un seuil donné. L'information sur les processus expliquant les phénomènes observés est inexistante. Selon leur compléxité, les modèles primaires sont caractérisés par un ou plusieurs paramètres comme le temps de latence, le taux de croissance la densité maximale etc. Ces paramètres sont spécifiques à des conditions d'environnement constant au cours du temps.
Les modèles décrivant l'effet des conditions d'environnement par exmple PH, température, acides sur les paramètres des modèles primaires sont appelés modèles secondaires ou mécanistes : ils sont construits à partir d'hypothèses expliquant les processus donnant lieu aux phénomènes observés. Les modèles utilisés pour faire le lien entre les modèles primaires et les modèles secondaires sont appelés modèles

tertiaires.

Le choix entre ces types de modèles est guidé par l'objectif donné à la modélisation . Si l'objectif est de comparer les caractéristiques de cinétiques bactérienne ou de prédire le taux de croissance ou de décroissance dans un environnement donné, un modèle empirique est suffisant. Si l'objectif de la modélisation est d'obtenir des informations sur le processus physiologique régulant le taux de croissance ou de décroissance, un modèle mécaniste est nécessaire.

En parallèle à cette classification, il a été defini des modèles dynamiques permettant de prédire la croissance microbienne en fonction des paramètres des modèles primaires qui varient au cours du temps.

2.2 Modélisation de la cinétique de croissance d'une population de micro-organismes

2.2.1 Différentes phases de la croissance

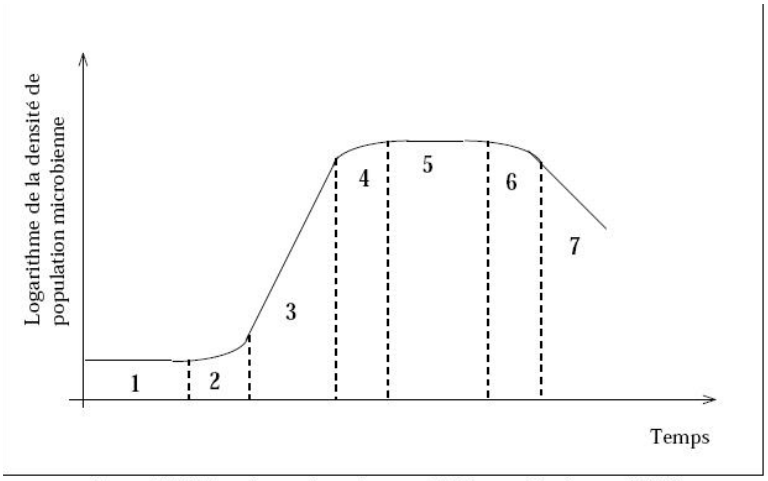


Figure II.1.3: Les phases de croissance établies par Buchanan (1918)

Le schéma standard de la croissance d'une population de micro-organismes en milieu non renouvelé a été établi par Buchnan (1918). Il décompose la cinétique d'une population microbienne en 7 phases :

1- phase initiale ou stationnaire ou phase de latence.
2- phase d'accélération de la croissance.
3- phase de croissance exponentielle.
4- phase de décélération de la croissance ou phase de freinage.
5- phase de stationnarité maximale.
6- phase d'accélération de la décroissance.
7- phase de décroissance exponentielle.

Il faut noter qu'en cinétique, il est difficile de représenter la croissance d'une population de micro-organismes en échelle arithmétique et le plus souvent on utilise l'échelle semi-logarithmique, et que les courbes observées dans la pratique s'écartent parfois de ce schéma. D'ailleurs cette décomposition n'est pas unique car d'autres microbiologistes ont proposé eux aussi des segments différents. Toujours est-il que les vraies courbes observées entraînent généralement un effort de modélisation.

2.3 Modélisation de la croissance

Le modèle de croissance le plus simple est le modèle exponentiel (*McMeekin et al.,*1993). Il suppose que la vitesse de variation de la densité de population microbienne x est proportionnelle à x, c'est à dire que $\frac{dx}{xdt} = \mu$ est constante.

$$\begin{cases} \frac{dx}{xdt} = \mu \\ t > 0\,,\ x > x_0 \geq 0 \end{cases}$$

Avec x la densité de la population à l'instant t, μ le taux de croissance maximale, x_0 densité de la population à l'instant $t = 0$, t le temps et dx est la variation de la densité bactérienne au cours d'un laps de temps dt. Avec la condition initiale x_0, la solution de l'équation différentielle est : $x_t = x_0 exp(\mu t)$ et elle est obtenue par la méthode de variation de la constante.
D'autres modèles existent comme les modèles de croissances décrivant les phases 4 et 5 à l'aide d'une fonction de freinage, et surtout en microbiologie alimentaire, les phases 1 à 5 de la croissance sont couramment décrites par le modèle de **Gompertz**, appliqué non pas à la densité de population microbienne, mais à son logarithme. Mais vu l'objectif assigné à la modélisation, nous allons utiliser le modèle de croissance exponentiel.

Comme on peut le constater sur la figure, la croissance apparaît souvent aprés une phase de croissance causé par un changement des conditions de croissance au moment de l'inoculation. Ce délai à la croissance est couramment désigné par le terme anglais *lag*. Le modèle défini précédemment n'intégre pas ce délai, ce qui fait que même si ce dernier ajuste bien la phase de croissance exponentielle, il est mal adapté pour décrire les croissances que l'on observe dans la nature.

2.4 Choix de modèle

On a choisi d'utiliser le modèle de croissance exponentiel avec temps de latence proposé par *Zamora et Zaritzky* en 1985 pour décrire l'évolution de la densité bactérienne x au cours du temps : il s'agit d'une extension simple du modèle exponentiel

permettant de tenir compte de la phase de latence.

$$\begin{cases} t < lag,\ x_t = x_0 + \varepsilon \\ t \geq lag\ , x_t = x_0 e^{\mu(t-lag)} + \varepsilon \\ \text{avec} \\ \varepsilon \rightsquigarrow N(0, \sigma^2) \\ \varepsilon_i\ \text{II}\ \varepsilon_j\ \forall\, i \neq j \end{cases} \tag{2.1}$$

En regardant un tel modèle, on peut avoir l'idée d'utiliser une transformation logarithmique pour avoir un modèle linéaire (le modèle linéaire est le plus maîtrisé par la communauté scientifique), cela répond certe à une demande de simplification, mais dans le cas des dénombrements sur boîtes de petri, la justification mathématique est pertinente.
En effet, plus la suspension bactérienne est diluée avant étalement, plus l'erreur associée au dénombrement est élevée. Au cours d'une croissance cette erreur augmente donc et la transformation logarithmique permet de réduire l'hétéroscédasticité des observations.
En ajustant ce modèle sur le logarithme base 10 on obtient : posons $X_t = log_{10}(x_t)$

$$\begin{cases} t < lag,\ X_t = X_0 + \varepsilon^* \\ t \geq lag\ \ , X_t = X_0 + \frac{\mu(t-lag)}{log(10)} + \varepsilon^* \end{cases} \tag{2.2}$$

Il évidement qu'on n'a pas la même modélisation de l'erreur.

Comme dans toute principe de modélisation et surtout en prédiction, il est intéressant de tenir compte de l'information disponible, et cela peut constituer une aide à la prédiction. En effet les travaux effectués par une équipe qui travaille sur la bactériologie à l'université de Lyon I et natamment les travaux effectués par Delignette et al (thèse pour l'obtention du diplome de doctorat université Lyon I, et l'article publié le 29 mars 1999, characterization of unexpected growth of Escherchia coli O157 H7 by Modeling) du laboratoire bactériologie, faculté de médecine Lyon et du laboratoire d'écologie microbienne et parasitaire de l'école nationale vétérinaire de Lyon, ont montré une corrélation biologique nette entre le temps de latence et le taux croissance μ. Les paramètres de croissance μ et lag sont inversement proportionnels. Ainsi la distribution du produit $\mu * lag$ semble être de type log-normale. La $v.a\ k = log(\mu * lag)$ a été approchée par une loi normale de moyenne m_k et σ_k. Cette information biologique peut être utile pour la qualité de la prédiction. Dans un souci de vouloir intégrer cette information dans le modèle, le modèle (2.2) a été reparamètré . Ainsi on :

$$\begin{cases} t < \frac{e^k}{\mu},\ X_t = X_0 + \varepsilon^* \\ t \geq \frac{e^k}{\mu},\ X_t = X_0 + \frac{(\mu t - e^k)}{log(10)} + \varepsilon^* \end{cases} \tag{2.3}$$

Comme on peut le remarquer, une part importante du travail dans le domaine de la modélisation de la croissance bactérienne en batch consiste en des ajustements

de divers modèles le plus souvent non linéaires à des jeux de données. Pour cette raison, il semble important de disposer de moyens pour pouvoir faire de la régression non linéaire. Dans la plupart des cas, on a à faire avec des modèles dynamiques : ce sont des modèles qui s'expriment en terme de population et qui décrivent l'évolution de celle-ci en fonction du temps.

2.5 Particularité de l'approche bayésienne

2.5.1 Modèle non linéaire

On considére un modèle non linaire de la forme :

$$y = f(\theta, x) + \varepsilon \tag{2.4}$$

Avec ε, l'erreur aléatoire additive distribuée normalement autour d'une moyenne nulle et de variance σ^2, y est la variable mesurée et x est la variable de contrôle associée.

L'approche bayésienne consiste à combiner deux sources d'information : l'information apportée par les données via l'expression d'une vraisemblance mais également une information à *priori* sur les paramètres, afin d'obtenir la distribution à *posteriori* des paramètres à estimer. L'inférence Bayésienne repose sur le théoreme de Bayes où la probabilté des paramètres conditionnellement aux données et notée $\mathbb{P}(Theorie/Donnee)$, est définie par :

$$\mathbb{P}(Theorie|Donnee) = \frac{\mathbb{P}(Theorie)\mathbb{P}(Donnee|Theorie)}{\mathbb{P}(Donnee)} \tag{2.5}$$

Avec $\mathbb{P}(Theorie)$, la probabilité à *priori* des paramètres, $\mathbb{P}(Donnee|Theorie)$, la fonction de vraisemblance des données conditionnellement aux paramètres, et $\mathbb{P}(Donnee)$, la probibilité totale des données obtenue en sommant le produit de la probabilité à *priori* et la fonction de vraisemblance sur toutes les valeurs des paramètres.

Un des avantages de cette approche est que l'information à *priori* peut permettre d'améliorer l'estimation de ces paramètres notamment en présence de nombreuses données manquantes, ces données étant traitées comme des paramètres inconnus.

2.6 Estimation des paramètres du modèle

Le théoreme de Bayes nous donne la densité de probabilité conditionnelle des paramètres connaissant les données :

$$\mathbb{P}(\theta, \sigma|y) = \frac{\mathbb{P}(\theta, \sigma, y)}{\mathbb{P}(y)} = \frac{\mathbb{P}(y|\theta, \sigma)\mathbb{P}(\theta, \sigma)}{\mathbb{P}(y)} \propto \mathbb{P}(y|\theta, \sigma)\mathbb{P}(\theta, \sigma) \tag{2.6}$$

C'est cette densité qu'on appelle densité à posteriori des paramètres, et l'inférence bayésienne joue un rôle important dans l'estimation du paramètre θ. Cette méthode a été appliquée à des modèles linéaires par *Lindley et Smith* en 1971, puis à modèles

non linéaires par *Berkey* en 1982 .

Soit $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ un $n-$échantillon des données. La vraisemblance des données peut s'écrire sous la forme :

$$\mathbb{P}(y|\theta, \sigma) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\theta, x_i))^2\} = \frac{1}{\sigma^n (2(\pi)^{\frac{n}{2}}} exp\{-\frac{1}{2\sigma^2} S(\theta)\} \tag{2.7}$$

$S(\theta)$ représente la somme des carrés des écarts .

Comme on dispose d'aucune information sur la loi à *priori* $\mathbb{P}(\theta, \sigma)$ des parmètres θ et σ, et en supposant l'indépendance de ces deux paramètres, on peut utiliser une distribution non informative sur σ, autrement dit on peut prendre $\mathbb{P}(\sigma) = \sigma^{-1}$ (cette distribution non informative a été également utilisée par *Box et Tiao* en 1973). En effet ce n'est pas une densité de probabilité car l'intégrale de cette quantité sur σ est infinie. L'utilisation d'une telle distribution à priori dite impropre est justifiée car l'intégrale de la densité à posteriori est finie sur l'espace des paramètres, et correspond donc à une normalisation prés à une densité de probabilité.

Ainsi connaissant la densité à priori des paramètres $\mathbb{P}(\theta)$, on peut écrire $\mathbb{P}(\theta, \sigma) = \sigma^{-1}\mathbb{P}(\theta)$. Par le théoreme de Bayes on obtient la distribution à posteriori des paramètres.

$$\mathbb{P}(\theta, \sigma|y) \propto \frac{\mathbb{P}(\theta)}{\sigma^{n+1}} exp\{-\frac{1}{2\sigma^2} S(\theta)\} \tag{2.8}$$

A partir de $\mathbb{P}(\theta, \sigma/y)$, on peut obtenir la distribution à posteriori marginale de θ par l'intégration sur l'écart type de σ :

$$\mathbb{P}(\theta|y) = \int_0^{+\infty} \mathbb{P}(\theta, \sigma|y) d\sigma \propto \mathbb{P}(\theta) S(\theta)^{-\frac{n}{2}} \tag{2.9}$$

On remarque ainsi que lorsque $\mathbb{P}(\theta)$ est constant ce qui équivaut à dire qu'on choisit pour θ une distribution localement uniforme le problème se raméne tout simplement à une minimisation de $S(\theta)$.

On peut tenir compte également de l'information biologique disponible et cela va nous permettre de choisir une loi à priori informative sur la $v.a\, k$. Ainsi en supposant toujours l'indépendance des résidus et en gardant la loi à priori non informative pour σ, on peut écrire :

$$\mathbb{P}(\mu, k, \sigma) = \mathbb{P}(\mu, \sigma)\mathbb{P}(k) \propto \sigma^{-1}\mathbb{P}(k) \tag{2.10}$$

Les paramètres de la loi k seront estimés dans les paragraphes suivants.
Puisqu'on avait supposé k suivant une loi normale, on peut écrire :

$$\mathbb{P}(k) = \frac{1}{\sigma_k \sqrt{2\pi}} exp(-\frac{1}{2}(\frac{k - m_k}{\sigma_k})^2) \tag{2.11}$$

Par l'hypothèse de normalité des résidus, la densité à posteriori des paramètres devient :

$$\mathbb{P}(\mu, k, \sigma|X) \propto \frac{1}{\sigma^{n+1}} exp(-\frac{S(\mu, k)}{2\sigma^2}) exp(-\frac{1}{2}(\frac{k - m_k}{\sigma_k})^2) \tag{2.12}$$

avec $S(\mu, k)$ la somme des carrés des écarts-type.

En intégrant cette densité à posteriori par rapport à σ on obtient la loi à posteriori marginale des paramètres μ et k

$$\mathbb{P}(\mu, k|X) = \int_0^{+\infty} \mathbb{P}(\mu, k, \sigma/X)d\sigma \propto S(\mu, k)^{-\frac{n}{2}} exp(-\frac{1}{2}(\frac{k - m_k}{\sigma_k})^2) \tag{2.13}$$

Une estimation des paramètres μ et k peut être obtenu par une par une maximisation de la densité à posteriori ce qui revient à une minimisation d'une fonction objective notée Θ et est definie par :

$$\Theta(\mu, k) = \frac{(k - m_k)^2}{2\sigma_k} + \frac{n}{2} ln(S(\mu, k)) \tag{2.14}$$

2.7 Principe de la regression non linéaire

Notre objectif est d'expliquer une variable aléatoire y continue (densité de la population microbienne) en fonction d'une variable temporelle t le temps en tenant compte des paramètres de croissance, on considére donc un modèle de la forme :

$$y_t = f(\theta, t) + \varepsilon_t \tag{2.15}$$

où f est une fonction mathématique choisie empiriquement sur une base de données observées, θ est le vecteur de paramètres inconnus à estimer, ε_t est le terme d'erreur du modèle qui exprime le caractère aléatoire (et donc non parfaitement prédictible) de y. De tels modèles sont dits dynamiques car décrivant la densité microbienne en fonction du temps.

Dans notre cas où le phénomène prend naturellement une forme non linéaire, l'estimation des paramètres devient nettement plus complexe que dans le cas de la régression linéaire classique car elle nécessite de recourir à des algorithmes itératifs souvent instables. Ainsi nous allons discuter des éléments de base sur l'estimation des paramètres d'un modèle non linéaire, et proposer une méthode numérique pour son ajustement.

2.7.1 Remarque

Le modèle (2.1) n'est pas linéarisable car on a supposé le terme d'erreur additif. Autrement dit l'expression proposée en (2.2) sera utilisée uniquement dans le but d'avoir une idée des paramètres initiaux pour l'algorithme itératif qu'on aura considéré.

Le modèle émanant de notre étude requiert toutes les bonnes propriétés un modèle non linéaire. En effet les modèles de croissance, les modèles Bayésienne et les équations différentielles sont généralement des modèles non linéaires. Comme dans le cas de la régression classique, nous allons utiliser la méthode des moindres carrés, qui est équivalente à celle du maximun de vraissemblance sous l'hypothèse supplementaire de la normalité des résidus, qui consiste à chercher les valeurs de θ qui minimisent la somme des carrés des résidus.

2.8 Méthode d'estimation

Pour chaque observation i $(i = 1, \cdots, N)$, le modèle à estimer s'exprime par l'équation suivante :

$$y_i = f(\theta, t_i) + \varepsilon_i \tag{2.16}$$

Avec $\mathbb{E}(\varepsilon_i) = 0$ et , $\mathbb{V}(\varepsilon_i) = \sigma^2$ constante.
Par une simple analogie de la régression linéaire, le critère à minimiser est

$$S(\theta) = \sum_{i=1}^{N} (y_i - f(\theta, t_i))^2 = (y - f(\theta))'(y - f(\theta)) \tag{2.17}$$

où $y = (y_1, \cdots, y_N)'$, et $f(\theta) = (f(\theta, t_1), \cdots, f(\theta, t_N))'$

Une technique pour minimiser le critère, consiste à le dériver par rapport aux paramètres et à chercher les solutions annulant ses dérivés. Le problème n'est pas aussi simple comme dans le cas linéaire où les dérivés forment un système de p équations linéaires à p inconnus en les paramètres : les équations normales. Dans le cas de la régression non linéaire, elles sont sous la forme :

$$min_\theta S(\theta) = min_\theta \sum_{i=1}^{N} (y_i - f(\theta, t_i))^2 \tag{2.18}$$

$$\Leftrightarrow \frac{\partial}{\partial \theta_j} S(\theta) = 0 \ \ \forall j = 1 \cdots p \Leftrightarrow \sum_{i=1}^{N} (y_i - f(\theta, t_i)) \frac{\partial}{\partial \theta_j} f(\theta, t_i) = 0 \ \ \forall j = 1 \cdots p \tag{2.19}$$

On peut les exprimer sous forme matricielle de la manière suivante :

$$F(\theta)'(y - f(\theta)) = 0$$

où $F(\theta)$ est la matrice $N * p$ dont l'élémnent (i, j) est la dérivée partielle $\frac{\partial f(\theta, t_i)}{\partial \theta_j}$

Ces équations pour la plupart du temps ont une forme très semblable au cas linéaire, mais malheureusement non linéaires en θ et ne peuvent se résoudre analytiquement : Il s'agit d'une non linéarité intrinséque.

Il existe plusieurs algorithmes itératifs, cependant dans ce rapport on a choisi de développer la méthode de **GAUSS-NEWTON**, cette dernière nous raméne dans le cas des modèles linéaires par développement de Taylor.

2.9 Méthode numérique de Gauss-Newton pour la résolution des équations normales

La méthode se base sur une linéairisation de la fonction $f(\theta, t)$. Elle peut s'approximer au premier ordre par un développement de **Taylor** autour d'un certain point θ^* :

$$f(\theta, t_i) \approx f(\theta^*, t_i) + \sum_{j=1}^{p} \left.\frac{\partial f(\theta, t_i)}{\partial \theta_j}\right|_{\theta=\theta^*} (\theta_j - \theta_j^*) \ \forall 1 \leq j \leq N$$

ou encore en notation matricielle

$$f(\theta) = f(\theta^*) + F(\theta^*)(\theta - \theta^*)$$

Si on remplace $f(\theta)$ par cette approximation dans $S(\theta)$, la fonction des moindres carrés à minimiser devient :

$S(\theta) = (y - f(\theta))'(y - f(\theta))$
$\cong (y - f(\theta^*) - F(\theta^*)(\theta - \theta^*))'(y - f(\theta^*) - F(\theta^*)(\theta - \theta^*))$
$= (y - f(\theta^*))'(y - f(\theta^*)) - 2(y - f(\theta^*))'F(\theta^*)(\theta - \theta^*) + (\theta - \theta^*)'F(\theta^*)'F(\theta^*)(\theta - \theta^*)$
$= S(\theta^*) - 2(y - f(\theta^*))'F(\theta^*)(\theta - \theta^*) + (\theta - \theta^*)F(\theta^*)'F(\theta^*)(\theta - \theta^*)$

$S(\theta)$ est donc approximée par une fonction de second ordre. Si on dérive cette expression de $S(\theta)$ et l'annule, on obtient :

$\frac{\partial}{\partial \theta} S(\theta) \cong \frac{\partial}{\partial \theta}(S(\theta^*) - 2(y - f(\theta^*))'F(\theta^*)(\theta - \theta^*) + (\theta - \theta^*)'F(\theta^*)'F(\theta^*)(\theta - \theta^*))$
$= -2F(\theta^*)'(y - f(\theta^*)) + 2F(\theta^*)'F(\theta^*)(\theta - \theta^*) = 0$

Le système d'équations résultant est linéaire en θ, et a pour solution :

$$\theta = \theta^* + (F(\theta^*)'F(\theta))^{-1}F(\theta^*)'(y - f(\theta^*))$$

Cette dernière équation va être utilisée comme base pour l'algorithme itératif de recherche de θ :

$$\hat{\theta}^{s+1} = \hat{\theta}^s + (F(\hat{\theta}^s)'F(\hat{\theta}^s))^{-1}F(\hat{\theta}^s)'(y - f(\hat{\theta}^s))$$

2.10 Propriétés des estimateurs

A ce niveau, il faut signaler que la taille de notre échantillon ne nous permet pas d'appliquer les propriétés asymptotiques des estimateurs. De ce fait, pour faire de l'inférence statistique sur le modèle (i.e tests sur θ ou prédictions) on a besion d'évoquer les propriétés suivantes :
$\hat{\theta}$ est estimateur par moindres carrés de θ ou par **M.L.E**
s^2 est estimateur de σ^2 defini par :

$$s^2 = \frac{1}{N-p} \sum_{i=1}^{N} (y_i - f(t_i, \hat{\theta}))^2$$

Alors on a :

$(\hat{\theta} - \theta) \rightsquigarrow N(0, \sigma^2(F(\theta)'F(\theta))^{-1})$

$\frac{(N-p)s^2}{\sigma^2} \rightsquigarrow \chi^2(N-p)$

$\frac{(\hat{\theta}-\theta)}{S(\hat{\theta})} \rightsquigarrow t(N-p)$

Ces propriétés génèrent moins d'erreurs de premières espèces que les résultats asymptotiques. Ces résultats s'utilisent pour tester si les paramètres de la régression sont significatifs ou pour calculer des intervalles de prédiction sur y. En particulier un intervalle de prédiction approché sur y en $t = t_k$ se calcule par la formule.

$$\hat{y}_k \pm t_{n-p}(1 - \tfrac{\alpha}{2})S(\hat{y}_k)$$

$$\Leftrightarrow f(t_k, \hat{\theta}) \pm t_{n-p}(1 - \tfrac{\alpha}{2})\sqrt{s^2 + s^2 \frac{\partial f(t_k,\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} (F(\hat{\theta})'F(\hat{\theta}))^{-1} \frac{\partial f(t_k,\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}}}$$

2.11 Problèmes numériques rencontrés dans l'utilisation des algorithmes itératifs

L'utilisation des algorithmes itératifs pour la résolutions de certains problèmes en mathématiques tels que l'estimation des paramètres d'un modèle non linéaire est souvent très delicate et peu parfois mener à des solutions incorrectes ou instables. Rien que pour un même problème, deux utilisateurs d'un même algorithme ou d'un même logiciel peuvent donner des solutions différentes. Raison pour laquelle on a évoqué le fondemment théorique de ces algorithmes pour se protèger de mauvais résultats.

2.12 Le choix du point de départ

Le point de départ joue un rôle important dans la convergence de l'algorithme. S'il est mal choisi, il peut s'arrêter avant d'avoir converger. C'est le critère d'arrêt qui décide quand l'algorithme s'arrête et la solution trouvée ne sera jamais le minimum exact de la fonction mais en sera théoriquement très proche. Différent critères d'arrêt sont possibles : On peut décider d'arrêter les itérations quand θ ne varie presque plus, ou quand la valeur du critère reste presque constant, ou encore aprés qu'un nombre maximale d'itérations ait attient. Il n'est donc pas étonnant de trouver des solutions différentes avec des logiciels différents. Cependant, dans autres situations, la fonction objective $S(\theta)$ à minimiser peut avoir plusieurs minima. Il s'agira alors de trouver entre ceux-ci le minimum absolu.

2.13 Ajustement des modèles sur les données et interprétation des coefficients

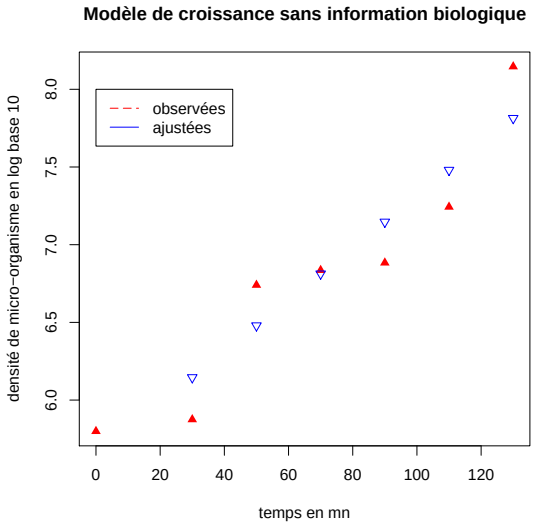
Le cas de l'Escherchia coli

Pour mettre en évidence le rôle important de l'information biologique recceuillie dans la littérature comme aide à la qualité de la prédiction, pour le cas de cet'espèce, on commence par ajuster le modèle exponentiel avec temps de latence sans tenir compte de cette information dont nous disposons (ie le modèle (2.2)). L'ajustement est fait par la technique de la régression non linéaire. Les paramètres de croissance correspondant à l'ajustement sont.

$\hat{\mu} = 0.0072$, **P-value** $= 0.0021$

$\hat{lag} = 49$, **P-value** $= 1.1610^{-6}$

En observant les données, on constate qu'il y'a une surestimation du paramètre lag, et la croissance semble s' effectuer de manière lente comme on peut le contater sur la figure.



Pour l'ajustement du modèle (2.3) (modèle avec information biologique), on peut envisager trois méthodes : Une première méthode consiste à estimer les paramètres μ et k en utilisant les techniques de la régression classique (MC2 : moindres carrés à deux paramètres). Aprés une estimation prélimaire des paramètres μ et k par une régression linéaire (pour avoir des valeurs initiales), on utilise la fonction **nls** (non linear regression) prédéfinie dans dans **R** et qui fait de la régression non linéaire par minimisation d'une fonction de perte. Elle utilise par défaut l'algorithme de **Gauss-Newton** dont le support théorique est annoncé plus haut, les résultats sont les suivants :

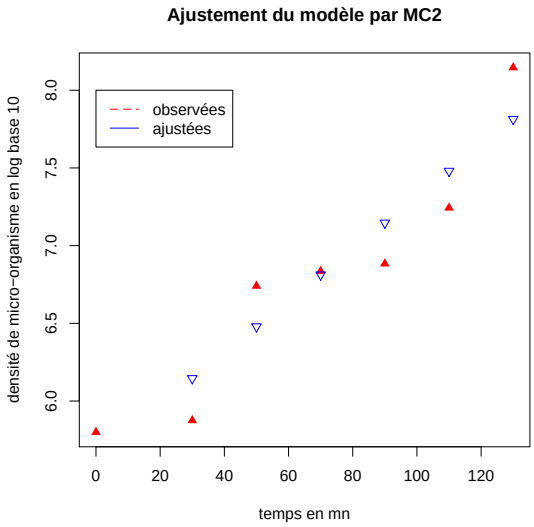
$\hat{\mu} = 0.038$, **P-value** $= 3.20610^{-4}$

$\hat{lag} = 9.24$, **P-value** $= 6.57710^{-5}$

$\hat{k} = -1.035$
La convergence est obtenue aprés 10 itérations avec une tolérance 1.12310^{-6} comme le témoigne la sortie logicielle

```
coli.res=nls(xt~(xi*exp(mu*(t1-lag))),
+      data=colli,start=list(mu=ini[1],lag=exp(ini[2])/ini[1]),trace=T)
coli.res$convInfo
$isConv
[1] TRUE

$finIter
[1] 10

$finTol
[1] 1.123725e-06

$stopCode
[1] 0

$stopMessage
[1] "convergence obtenue"
```

x_t est la variable contenant les tailles des inocula, x_i correspond à la densité obtenue à $t = 0$, $ini[1]$ et $ini[2]$ contiennent les valeurs intiales estmées directement par la regression linéaire (la fonction fournissant ces valeurs initiales n'est pas prédéfinie dans **R**). Le code 0 est l'équivalent du message d'arrêt obtenu, et le tout a été obtenu par appel à la fonction contenant les informations de la convergence (convInfo). Ainsi on constate qu'en intégrant l'information biologique dans la modélisation la capacité prédictive du modèle s'est améliorée, et ceci peut s'expliquer par une valeur du temps de latence devenue beaucoup plus raisonnable car d'aprés les données il se situe entre 0 et 30 mn et son test de significativité (**P-value**) montre tout son importance dans la cinétique.

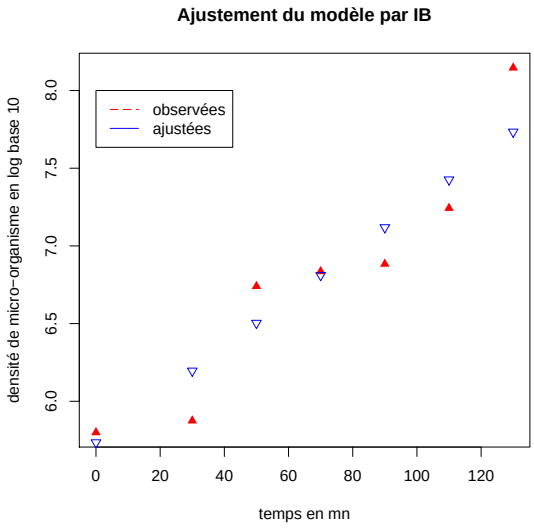


La deuxième méthode consiste tout simplement à réduire le nombre de paramètres à estimer en fixant k égale à une constante. En effet, cette dernière méthode fait partie des techniques d'inférences bayésienne. On peut refuser toute information à priori sur les paramètres, et dans ce cas les techniques d'inférence bayésienne prennent en charge notre ignorance sur les paramètres. Dans ce cas, il faut choisir une loi à priori non informative $\pi \propto 1$, évidemment d'autres possibiltés existent mais celle-ci est la plus simple. Dans ce travail, cette méthode n'a pas été implémentée car elle peut se déduire de la troisième en posant $k = m_k$ (voir l'expression (2.14)). En faite, c'est une technique d'inférence bayésienne (IB) vue dans un angle particulier.
La troisième méthode consiste à ajuster le modèle de croissance en utilisant la fonction objective obtenue par l'approche bayésienne. Pour caractériser l'information à priori sur le paramètre k, on prend $m_k = -0.08$ et $\sigma_k = 0.26$ (ce choix sera justifié ultérieurement). Avec cette méthode, on a utilisé la fonction prédéfinie dans **R** **nlm** (non linear minimisation) qui minimise des fonctions non linéaires par approche itérative utilisant le seul algorithme **Newton**. La valeur minimale de la fonction est obtenue aprés 16 itérations avec $\hat{\mu} = 0.03541$, $\hat{k} = -1.899$ et une valeur de $\hat{lag} = 4.223$. Elle fournie les plus petites valeurs des paramètres de croissance par rapport autres méthodes. Aprés l'appel de la fonction on a obtenu la sortie suivante.

```
nl=nlm(fnlm,c(0.03,-1.9),hessian=T)
$minimum
[1] -23.28909

$estimate
[1]  0.03541358 -1.89997932

$gradient
[1] -6892902.23     6550.94

$hessian
     [,1]     [,2]
```

```
[1,] -Inf        0
[2,]    0 -8847472

$code
[1] 2

$iterations
[1] 16
```

fnlm est la fonction objective à minimiser, 0.03 et -1.9 sont des paramètres initiaux à fournir l'algorithme car elle est itérative. La composante estimate donne les valeurs de $\hat{\mu}$ et $\hat{k}$ correspondant au minimum de la fonction qui dans ce cas précis vaut -23.28909. Dans l'aide du logiciel **R**, la phrase suivante est l'équivalent du code d'arrêt 2 indiquant que notre algorithme a convergé :
"successive iterates within tolerance, current iterate is probabily solution".



En comparant les deux ajustements (MC2 et IB) on voit nettemment que l'ajustement fourni par la méthode IB est meilleur, non seulement elle s'approche mieux la cinétique, elle ajuste très bien le premier point de la cinétique, tandis qu'avec la méthode MC2 ce point n'est pas ajusté et il est important car correspondant à la première mesure précoce de la cinétique (voir figures ajustement avec MC2 et ajustement avec IB). Il n'est donc pas possible à partir des premiers points de la mesure d'ajuster le modèle par la méthode MC2. La méthode IB permet ainsi d'avoir une prédiction raisonnable de la cinétique à partir de ses premiers points expérimentaux.

Le cas Klesbsiella pneumoniae

Avec les jeux de données qui suivent seulement les modèles intégrant l'information biologique ont été ajustés car ce sont les seuls pouvant donner les résultats appréciables. Les paramètres de croissance qui ont été estimés par la méthode MC2 valent :

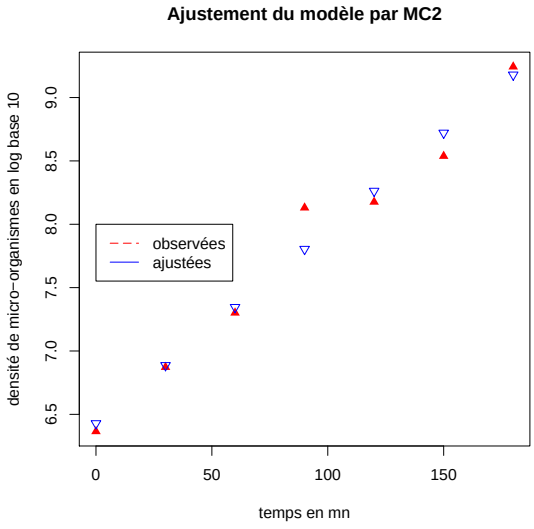
$\hat{\mu} = 0.035$, **P-value** $= 1.610^{-5}$

$\hat{lag} = 29$, **P-value** $= 0.015$

$\hat{k} = 0.022$

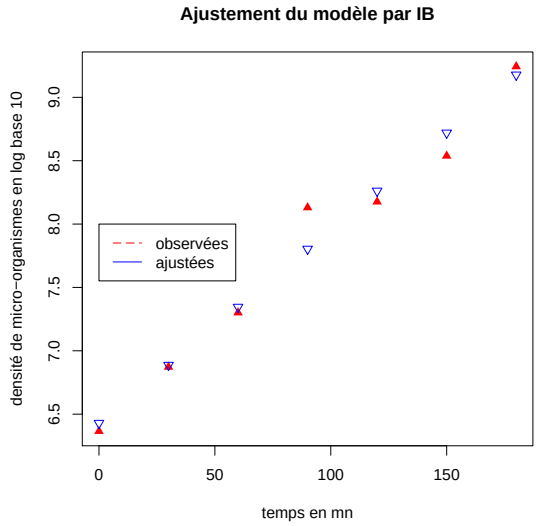
La convergence est obtenue aprés 7 itérations avec une tolérance 9.210^{-6}. La sortie logicielle correspondante est

```
> sella.res=nls(xt~(xi*exp(mu*(t1-lag))),
+       data=seela,start=list(mu=ini[1],lag=exp(ini[2])/ini[1]),trace=T)
> sella.res$convInfo
$isConv
[1] TRUE

$finIter
[1] 7

$finTol
[1] 9.206341e-06

$stopCode
[1] 0

$stopMessage
[1] "convergence obtenue"
```



Avec l'ajustement par minimisation de la fonction objective ou par critère bayésienne, le temps de latence et le taux de croissance restent pratiquement inchangés (voir sortie logicielle). Il faut noter que même si on trouve pratiquement les même valeurs, avec cette méthodes le nombre d'itérations est beaucoup plus imporatant. La méthode MC2 fait 7 itérations tandis que la méthode IB en fait 19 pour trouver des résultats dont la différence n'est pas significative.

```
> nl=nlm(fnlm,c(0.03,0.02),hessian=T)
$minimum
[1] -21.63355

$estimate
[1] 0.03516046 0.01993981

$gradient
[1] -499917.890    5276.844

$hessian
     [,1]     [,2]
[1,] -Inf        0
[2,]    0 -6063510

$code
[1] 2

$iterations
[1] 19
```



Dans le cas des données de l'espèce Escherchia coli, la cinétique a démaré avec un inoculum de 6.310^5 et la méthode IB s'est montrée beaucoup plus performante pour prédire la courbe de la croissance. Pour le cas de Klesbseilla pneumonia, la cinétique a débuté à partir de 2.3210^6 et les résultats fournis par les deux ajustements sont pratiquement identiques. Cependant on peut même dire que la méthode MC2 s'est révélée plus performante que la méthode IB car celle-ci fait moins d'itérations, donc préférable du point de vue compléxité algorithmique.

2.14 Analyses des résultats

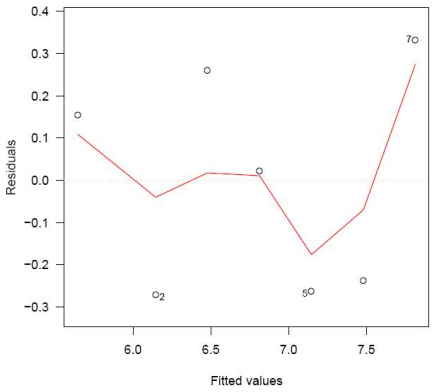
Pour une prédiction précoce de la croissance, la méthode IB donne une meilleur précision que la méthode MC2. Cependant, pour des densités de populations d'envoron 10^6 les méthodes MC2 et IB atteignent quasiment la même précision.
Il faut noter que pour les inocula faibles, la méthode MC2 n'a pas reuissi à donner un bon ajustement du début de la cinétique, certainemment l'inforamation donnée par les premiers points de la mesure est insuffisante. Dans ce cas , la méthode IB permet de donner une prédiction raisonnable de la courbe de croissance.

Les résultats précédents ont été obtenus en prennant comme paramètres de distribution de k égalent à -0.08 et 0.26 . En effet dans l'article où cette informtion biologie a été obtenue les auteurs ont choisi comme paramètres de k la moyenne et l'écart-type de trois valeurs de k obtenues sur trois espèces. Ignorant le pourquoi de ce choix, on a préféré estimer les paramètres de la distibution de la variable aléatoire k par les techniques du **Boostrap**. Cette dernière technique nous permet d'estimer des quantités sur des données dont on ignore la loi qui les génére : on part du principe qu'on dispose uniquement les données mais pas autre information.
Toujours est -il qu'en pratique, lorsque l'on veut faire une prédiction, il est très rare que l'on puisse disposer d'une telle connaissance, raison pour laquelle il est intéressant de savoir la répercussion des paramètres de cette loi sur la performance des méthodes de prédiction de la cinétique. Notons aussi qu' une bonne courbe de croissance suppose une bonne estimation du *lag* et qu'il se trouve dans les conditions optimales de croissance, il devient de plus en plus petit ce qui rend difficile son estimation.

2.15 Examen des hypothèses sur le modèle d'erreur

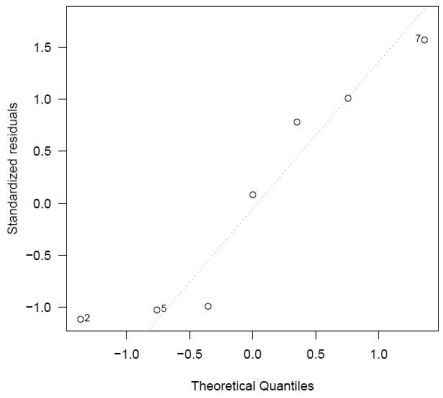
Afin de vérifier que la considération d'un modèle d'erreur addtif indépendant normalement distribué autour d'une moyenne nulle implicite à la méthode des moindres carrés n'est pas aberrante, nous avons examiné les graphiques des résidus en fonction des valeurs prédites.

Résidus indépendants et centrés



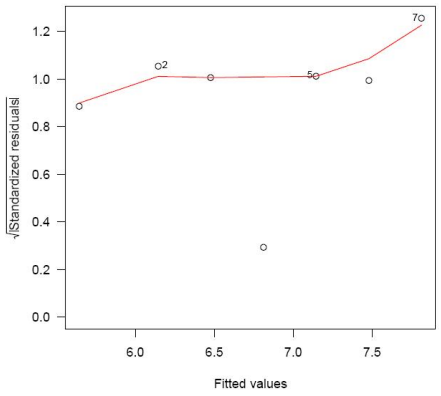
Aucune forme particulière de ce graphe pouvant remettre en cause l'hypothèse d'un modèle d'erreur nulle ou l'indépendance n'est pas mise en évidence.

Normalité des erreurs



Le grahique des quantiles normales en fonction des quantiles théoriques ne contredit pas l'hypothèse de la normalité des résidus.

Homoscédasticité des erreurs



Aucune variation évidente de l'amplitude des résidus n'ayant pas été obsevée au cours de la croissance.

Chapitre 3

Modélisation des caractères biochimiques en fonction du temps d'incubation et de la taille des inocula

3.0.1 L'échec de la regression logistique

Cette partie concerne les deux problèmes les plus couramment rencontrés en biologie, à savoir évaluer le potentiel relatif d'une substance inconnue par rapport à une substance standard, estimer la réponse à un stimulus. Le plus souvent ce problème est résolu par la régression logistique binaire car dans la majorité des cas, la réponse au stimulus est de type "tout ou rien" (positive ou négative).
Une fois de plus, il est intéressant de bien mettre en exergue, le type d'analyse qui concerne notre étude. En effet, selon les connaissances biomédicales, il faut distinguer entre une analyse "pronostique" et une analyse "étiologique" car le choix d'une des ces analyses a un impact sur la finalité du modèle à choisir. Dans une analyse étiologique, on s'intéresse particulièrement à évaluer le risque associé à un facteur et le choix des facteurs confondants est primordial pour éliminner les biais autant que possible, tandis que pour une analyse "pronostique", on cherche avant tout à construire un modèle permettant de prédire (on dit discriminer en terme de régression logistique) le mieux possible les **outmes** (tout ou rien) à partir des covariables. Cette partie concerne essentiellement ce type d'analyse.

On rappelle que l'étude porte sur quatre espèces que sont : Le Escherichia coli, le Klesbsialla pneumonia, Sallmonella paratyphi A et Acinetobacter.
On a n tests biochimiques, I inocula et J temps d'incubation. Pour chaque souche, pour chaque inoculum i fixé, on cherche les réponses des n marqueurs pour chaque temps d'incubation $t_j \in \{4, 8, 12, 24\}$. Certainemment les valeurs 4 et 24 ont été choisies pour faire un parallèlisme avec les galeries API. En effet pour ces galeries commercialisées, le temps d'incubation minimale et maximale sont respectivement 4 et 24.

3.0.2 Le cas Escherichia coli

On a des souches ensemencées au temps $t = 0$ et avec une certaine quantité d'inoculum qu'on suit à différent temps. Pour chacun des temps, on regarde en fonction du nombre de colonies le dégre de concordance des caractères biochimiques avec une certaine reférence pour l'Escherchia coli. Aucune indication sur le nombre d'expériences par "cases" du tableaux, encore moins de la variabilité de la réponse n'a été mentionnée par les tableaux des données.

Remarque ici la concordance a été définie comme étant le pourcentage de carcactères biochimiques justes par rapport à la reférence. En terme épidémiologique c'est ce qu'on appelle la prévalence.

Dans la mesure où on a effectivement une reférence qui joue un rôle privilégier, il convient d'arriver à caractériser la probabilité du n-uplet des marqueurs.
$\mathbb{P}(M_1 = 1, M_2 = 1, \cdots, M_n = 1|$taille inoculum, temps d'incubation, autres facteurs de confusions possible).
Si on considére les tests indépendants au moins conditionnellement au temps et à la taille de l'inoculum, on peut écrire la probabilité globale en fonction du temps d'une concordance avec la signature de l'espèce :
$\mathbb{P}(M_1 = 1, M_2 = 1, \cdots, M_n = 1$ |taille inoculum, temps d'incubation)
$= \prod_{i=1}^{n} \mathbb{P}(M_i = 1|I_k, t_j)$, avec $\mathbb{P}(M_i = 1|I_k, t_j)$ la probabilité que le marqueur M_i répond positif au temps d'incubation t_j avec un inoculum I_k
$\prod_{i=1}^{n} \mathbb{P}(M_i = 1|I_k, t_j) = \prod_{i=1}^{n} B(p)$, avec $p = f(I_k, t_j)$ i.e chaque marqueur M_i suit une loi de bernouille de paramètre p, et donc le produit suit une loi binomiale de paramètres n et p
$\prod_{i=1}^{n} \mathbb{P}(M_i = 1|I_k, t_j) = B(n, p)$

Remarque Dans le cas où on ne suppose pas l'indépendance des tests, le problème est tout a fait delicat. En effet, il faudrait faire intervenir les corrélations $\rho(ij)(t)$ entre les réponses pour 2 marqueurs i et j en faisant des hypothèses sur l'évolution au cours du temps de ces corrélations, et cela nous laisserait dans la situation des tests binaires multiples imparfaits (i.e avec une certaine sensibilité et une certaine spécificité) et corrélés.

3.0.3 Définitions

La sensibilité est définie comme la probabilité de classer l'individu dans la catégorie $y = 1$ (on dit que le test est positif), étant donné qu'il est effectivement observé dans celle-ci :
sensibilité= $\mathbb{P}(test + |y = 1)$

La spécificité est défini comme étant la probabilité de classer l'individu dans la catégorie $y = 0$ (on dit que le test est négatif), étant donné qu'il est effectivement observé dans celle-ci :
specificité= $\mathbb{P}(test - |y = 0)$

Ainsi sous l'hypothèse d'indépendance des tests, caractérisé chacun par une cer-

taine sensibilité et une certaine spécificité, et tenant compte de la prévalence de la souche considerée dans la population, on peut écrire la probabilité d'avoir observé un profil particulier de réponse U en fonction de l'inoculum et du temps d'incubation, et un moyen pour y parvenir parmi tant d'autres peut être le modèle **logit**. Le but était de proposer un modèle logit pour chaque test et ou un modèle logit global pour U, car la réponse de U ne saurait être que la combinaison des réponses partielles des différents tests.

Raison pour laquelle pour prouver l'échec de ce modèle de manière modélisatrice, on a considéré les variables les plus "intéressantes"(les tests qui durant l'incubation ont été positifs et négatifs) dans les tableaux des données pour les modéliser avec la fonction logit, et voici une sortie logicielle.

```
maglm=glm(man~i+t,family=binomial,data=data2)
Warning message:
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,:
des probabilités ont été ajustées numériquement à 0 ou 1
summary(maglm)

Call:
glm(formula = man ~ i + t, family = binomial, data = data2)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-5.847e-06  -1.446e-06   2.107e-08   1.970e-06   7.816e-06

Coefficients:
              Estimate Std. Error   z value Pr(>|z|)
(Intercept) -1.228e+02  5.877e+05 -2.09e-04        1
i           -7.216e-11  7.601e+04 -9.49e-16        1
t            7.070e+01  1.182e+05     0.001        1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2.2493e+01  on 19  degrees of freedom
Residual deviance: 4.7641e-10  on 17  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```

Dans la variable man (voir tableaux des données) on a utilisé le codage suivant : 0 si la réponse du test est négatif, 1 s'il est positif. i est la variable contenant les densités des souches, et t contient les temps d'incubations, l'option family indique ici la fonction logistique utilisée et la distribution de l'erreur qui en faite dans ce cas précis suit une loi de bernouille, et glm est la fonction prédéfinie dans **R** qui

fait de la régression logistique. En effet, **R** nous dit que la totalité de l'information cherchée est contenue dans la variable t, c'est comme si la variable i ne contient aucune information pour man : les résidus sont tous nuls (les probabilités ont été ajustées numériquement à 0 ou à 1), et les tests **Wald** sont hautements non significatifs ($Pr(>|z|)$: ils sont tous égalent à 1.
Cependant, si l'on regarde bien les données ce résultat n'est pas surprennant car si $t = 4$ man=0 et man vaut 1 si t différent de 4 et quel que soit l'inoculum considéré. Avec des souches différentes et des exemples de **R** a l'appui avec des variables différentes ont donné des résultats similaires.

Cette démarche a été testée sur des données fictives (tirées à partir des données dont nous disposons par une simple régles de trois aux temps $t = 5, 6, 7, ...12$, en effet ceci peut ne pas être aberrant si on est réellement dans la phase exponentielle de croissance), et les résultats sont prometteurs. Cette démarche fait partie des méthodes candidates pour donner une analyse pertinente du temps d'incubation si l'on dispose suffisamment de mesures dans la phase exponentielle de croissance et ceci de manière séquentielle aux temps $t = 5, 6, 7, ...$ par inoculum.
Etant certain que la méthode de la régression logistique ne répond pas à nos attentes on ne peut être à l'abri de la recherche d'autres alternatives. La première idée mise en avant était de modéliser le temps d'incubation en fonction de l'inoculum et de la concordance des caractères biochimiques. L'objectif était de fixer notre inoculum et une quantité de concordance raisonnable à une identification acceptable de l'espèce et de chercher le temps d'incubation qui en resulte. Aprés plusieurs tentatives, tous les modèles (linéaires comme non linéaires) identifiés ont laissé hautement non significatif l'inoculum, avec des **P-values** minimales atteignant les 0.1, et ceci n'est pas pertinemment vue les résultats des expériences.
Une autre tentative à consister a lire la quantité de la concordance des caractères biochimiques en fonction de l'inoculum et du temps d'incubation.

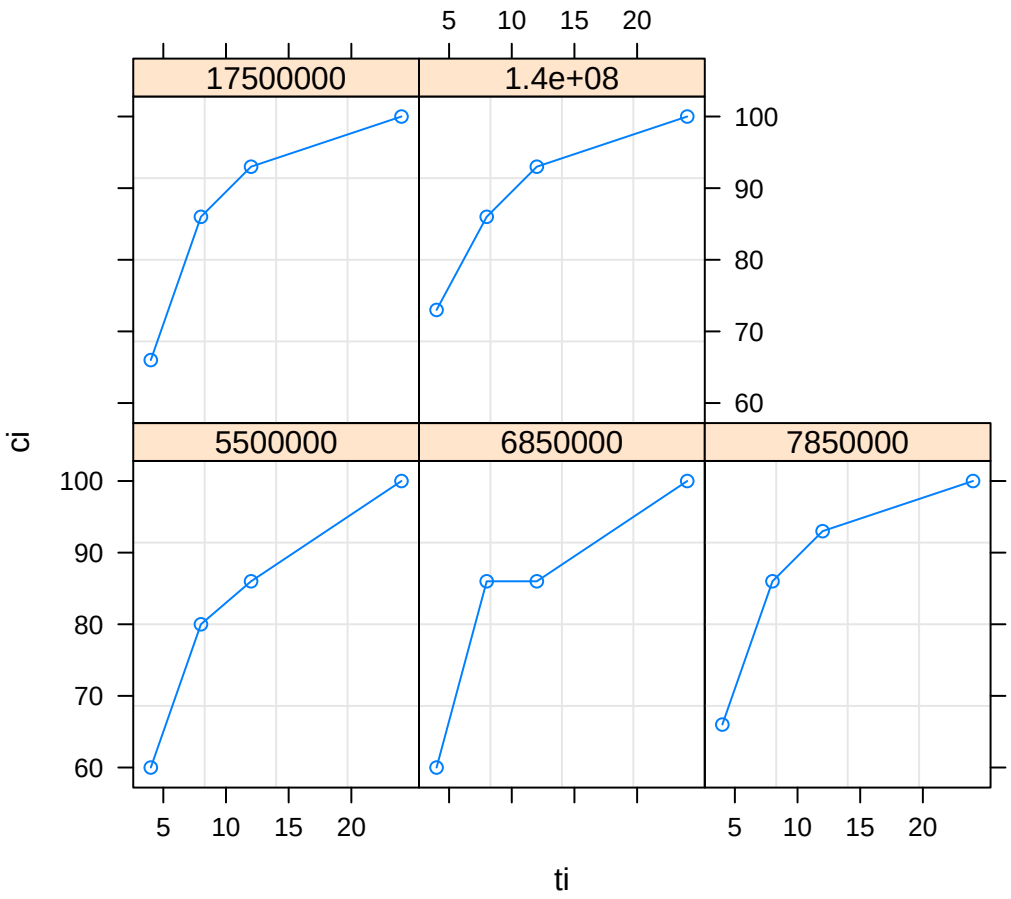
3.1 Quantification de la concordance des caractères biochimiques

En faisant une analyse des tableaux de données, la variable concordance des caractères biochimiques fait apparaître deux stratégie de comparaison : Une stratégie de comparaison globale (données répetées et correlées, cette dernière fera l'objet d'une autre contribution personnelle en fin de chapitre), et une stratégie de comparaison séquentielle. En d'autres termes, on ne regarde la concordance à 8 h que s'il y'a absence de discordance à 4 h, et idem à 12 h en fonction des résultats à 4 h, 8 h, etc. Toute tentative de modélisation nécessite une étude descriptive préalable afin de s'assurer au moins graphiquement de la validité des hypothèses considérées. On propose une représentation graphique du nuage de points et une régression nonparamètrique afin de déceler une éventuelle liaison non linéaire entre les variables.

3.1.1 Analyse grahique des données

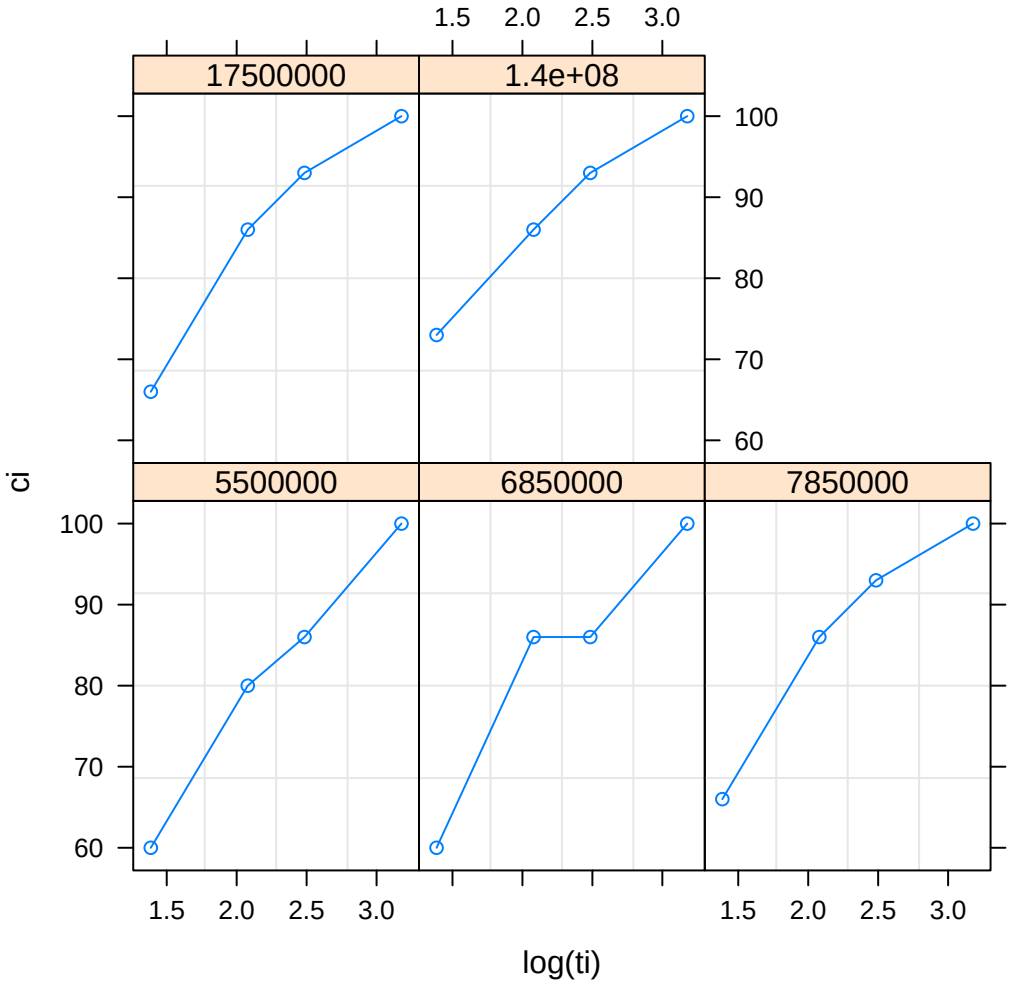
Le cas de l'Escherchia coli

La variable *ci* désigne les concordances de l'espèce i et *ti* la variable des temps d'incubation.



On constate sur la figure que si l'inoculum est fixé, la concordance des caractères biochimiques croit de manière logarithmique en fonction du temps d'incubation. On peut proposer le modèle de regression non-paramètrique de la forme :
$c_i^{t_j} = f(t_j, I_i) + \varepsilon_{ij}(t)$, avec $t_j \in \{4, 8, 12, 24\}$, $i = 1, ..., n$, f fonction croissante avec $\varepsilon_{ij}(t)$ corrélées pour chaque couple (i, j) donné.
$c_i^{t_j}$ représente le dégré de concordance obtenu pour un inoculum i fixé et un temps d'incubation $t_j \in \{4, 8, 12, 24\}$.
D'aprés la représentation graphique précédente, des problèmes de non linéarité sont identifés, l'étape suivante consiste à rechercher des transformations élémentaires susceptibles de les résoudre. Dans ce cas la transformation logarithmique semble bien

adapter. Ceci améne à faire la représentation graphique en fonction du logarithme de variable temporelle.



Ainsi avec cette tansformation, on constate que la considération d'un modèle linéaire entre la concordance des caractères biochimiques en fonction du logarithme du temps par inoculum fixé n'est pas absurde, et la principale variable d'intérêt est le temps.

3.1.2 Approche modélisatrice

En mathématique, l'analyse des résultats d'une expérience relève de l'emploi d'un modèle de régression lorsque chaque observation de la variable réponse peut être représentée comme la somme d'un terme dépendant de la valeur prise par une variable contrôle ou plusieurs et de la réalisation d'une variable aléatoire (erreur). On établi un modèle régression linéaire multiple en expliquant la concordance à l'aide des variables t et i. On avait signalé précédemment qu'on n'avait aucune idée du nombre d'expériences par "case" du tableau, pour des raisons de simplicité on

prend $n_{ij} = 1$, et il n'est pas possible d'ajouter un terme d'intéraction entre les deux variables. Cette analyse a été améliorée en utilisant les transformations log_{10} pour inoculum et log pour le temps. Aprés toutes les tentatives, on a identifié le modèle suivant :

$$c_i^{t_j} = \mu + \alpha_i i_i + \beta_j t_j + \varepsilon_{ij}$$

α_i mesure la contribution de l'inoculum i_i dans la concordance des caractères biochimiques.
β_j mesure la contribution du temps d'incubation t_j dans la concordance des caractères biochimiques.

En notation vectorielle, on peut écrire pour n observations :

$$C = X\theta + \varepsilon$$

Avec C : le vecteur de dimension n des concordances ; $X\theta$: le vecteur de composantes $f(t_j, I_i)$, $i = 1, \cdots, n$; θ : le vecteurs des paramètres ; ε : le vecteur centré de dimension n des erreurs et X : la matrice des inocula et des temps d'incubation.

L'analyse d'un modèle de régression linéaire dépend des hypothèses suivantes. Le vecteurs θ des "vrais" paramètres inconnus appartient à un ensemble Θ (espace des paramètres) d'intérieur non vide. Pour toute valeur fixée des régresseurs, la fonction $X\theta$ est injective et deux fois continûment dérivable en tout point intérieur à Θ. Les termes d'erreurs sont indépendants et identiquement distribués de moyenne nulle et de variance constante σ^2. Cette dernière hypothèse modélisant l'erreur permet de rendre compte de la nature aléatoire des phénomènes étudiés. L'objet de la régression est donc de modéliser et d'étudier sous tous les aspects la relation entre la variable réponse (concordance des caractères biochimiques) et les variables explicatives (inocula et temps d'incubation).

3.2 Rappel sur les modèles linéaires

Le modèle est linéaire et s'exprime comme somme des paramètres mulitipliés par les variables. Il se représente par un plan dans un espace à $n + 1$ dimensions. L'estimation est le choix d'une estimation de θ notée $\hat{\theta}$ dans l'ensemble Θ compte tenu des observations, c'est à dire des résultats des expériences. Pour choisir un estimateur, on utilise la méthode des moindres carrés : elle consiste à minimiser la distance entre le le modèle et les observations. Cette notion de distance pose en général un problème de choix de la métrique. Le critère le plus courant est géométrique et a été proposé par *Laplace*. Il consite à choisir un vecteur θ telle que la norme euclidienne entre le vecteur des concordances prédites et le vecteur des concordances observées soit minimale. C'est donc un vecteur θ qui minimise la fonction de perte définie par : $||C - X\hat{\theta}||^2$. Elle est équivalent à la somme des carrés des écarts résiduels.
$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Sous l'hypothèse d'indépendance, et d'homogénéité de variance des résidus, le meilleur estimateur de θ par la méthode des moindres carrés est :

$$\hat{\theta} = (X'X)^{-1}X'C$$

La matrice du développement peut s'écrire sous la forme :

$$\mathbf{X} = \begin{pmatrix} 1 & i_1 & t_1 \\ 1 & i_1 & t_2 \\ 1 & i_1 & t_3 \\ 1 & i_1 & t_4 \\ 1 & i_2 & t_1 \\ 1 & i_2 & t_2 \\ 1 & i_2 & t_3 \\ 1 & i_2 & t_4 \\ \vdots & \vdots & \vdots \\ 1 & i_n & t_1 \\ 1 & i_n & t_2 \\ 1 & i_n & t_3 \\ 1 & i_n & t_4 \end{pmatrix}$$

La matrice des paramètres peut s'écrire sous la forme :

$$\theta = \begin{pmatrix} \mu & \alpha_1 & \beta_1 \\ \mu & \alpha_1 & \beta_2 \\ \mu & \alpha_1 & \beta_3 \\ \mu & \alpha_1 & \beta_4 \\ \mu & \alpha_2 & \beta_1 \\ \mu & \alpha_2 & \beta_2 \\ \mu & \alpha_2 & \beta_3 \\ \mu & \alpha_2 & \beta_4 \\ \vdots & \vdots & \vdots \\ \mu & \alpha_n & \beta_1 \\ \mu & \alpha_n & \beta_2 \\ \mu & \alpha_n & \beta_3 \\ \mu & \alpha_n & \beta_4 \end{pmatrix}$$

Le vecteurs des erreurs peut s'écrire lui aussi sous la forme :

$\varepsilon = (\varepsilon_{11}, \varepsilon_{12}, \varepsilon_{13}, \varepsilon_{14}, \varepsilon_{21}, \varepsilon_{22}, \varepsilon_{23}, \varepsilon_{24}, \cdots, \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3}, \varepsilon_{n4})'$

3.2.1 Ajustement du modèle sur les données et interprétation des coefficients

Les données de l'Escherchia coli

On ne va pas tenir compte dans nos interprétations du terme intercept (μ) (qui modélise l'effet général) qui d'ailleurs s'est revélé hautement non significatif pour tous les jeux de données où le modèle a été ajusté. Pour le cas de E.coli par exemple son test de nuleté vaut $P - value = 0.26$, l'heureustique de la modélisation veut que même si ce dernier est non significatif qu'il ne soit pas supprimé du modèle. Aprés ajustement du modèle sur les données de E.coli, on obtient les statistiques estimées :

$\hat{\alpha} = 3.645$, $\hat{\beta} = 19.231$, $\mathbf{R}^2 = 0.92$, $\mathbf{R}^2_{ajuste} = 0.91$, et $\mathbf{F} = 4.18210^{-10}$

Pour avoir une idée de la pertinence de chaque paramètre, on utilise le test de **student**. En effet ce test nous permet de décider parmi les hypothèses suivantes laquelle faut-il rejetter.

$$\begin{cases} H_0 & \alpha = 0 \\ \text{vs} & \\ H_1 & \alpha \neq 0 \end{cases}$$

Pour décider parmi ces deux hypothèses laquelle est juste, on utilise la **P-value**. En effet pour le paramètre α, cette dernière statistique vaut $\mathbf{P} = 0.0512$, et on voit qu'au seuil 0.05 on n'est pas loin d'accepter l'hypothèse H_0 , tandis que pour β on :

$$\begin{cases} H_0 & \beta = 0 \\ \text{vs} & \\ H_1 & \beta \neq 0 \end{cases}$$

le test est hautement non significatif avec une **P-value** de 9.8610^{-11}, ce qui entraine le rejet automatique de l'hypothèse nulle.

Dans notre souci de faire une identification avec un temps d'incubation inférieur à 24 h, et en guise d'amélioration du modèle, on décide de soustraire des données toutes les informations relatives à 24 h

En ajustant à nouveau le modèle sur les données dont on a enlevé les colonnes de 24 h, les statistiques sont de nouveau estimées et elles valent :

$\hat{\alpha} = 4.86$, $\hat{\beta} = 23.544$, $\mathbf{R}^2 = 0.91$, $\mathbf{R}^2_{ajuste} = 0.90$, et $\mathbf{F} = 2.80910^{-7}$

Comme on le constate les valeurs des paramètres ont changé, mais ce qu'on a gagné de plus est la précision du paramètre α. En effet les **p-values** ont changé et on trouve 8.810^{-8} pour β et 0.02 pour α toujours au seuil de 0.05. L'information de 24 h peut être considérée comme une information certaine (une information évidente) car quel que soit l'inoculum qu'on considére, une fois incubé jusqu'à 24 h, l'identification de l'espèce sera possible.
Quant à la statisque $\mathbf{R}^2$, c'est ce qu'on appelle coefficient de détermination, elle fait partie des statistiques qui quantifient la qualité de l'ajustement du modèle sur les données (goodness of fit). Le coefficient de détermination est une mesure d'association qui tient qui se rattache à la famille des mesures de similarité, plus exactement, c'est une mesure du degré d'accord entre le modèle et les observations : il exprime la part de la variabilté "expliquée" dans la variabilité totale. C'est un critère quantitatif qui permet de comparer différent modèles ayant le même nombre de variables. A nombre de paramètres égal, et sous réserve du non-rejet des hypothèses sur le modèle d'erreur, le modèle dont le coefficient de détermination le plus élevé assure le meilleur ajustement. Raison pour laquelle ce dernier a nécessité un ajustement et c'est le coefficient $\mathbf{R}^2_{ajuste}$ qu'est utilisé dans la pratique.

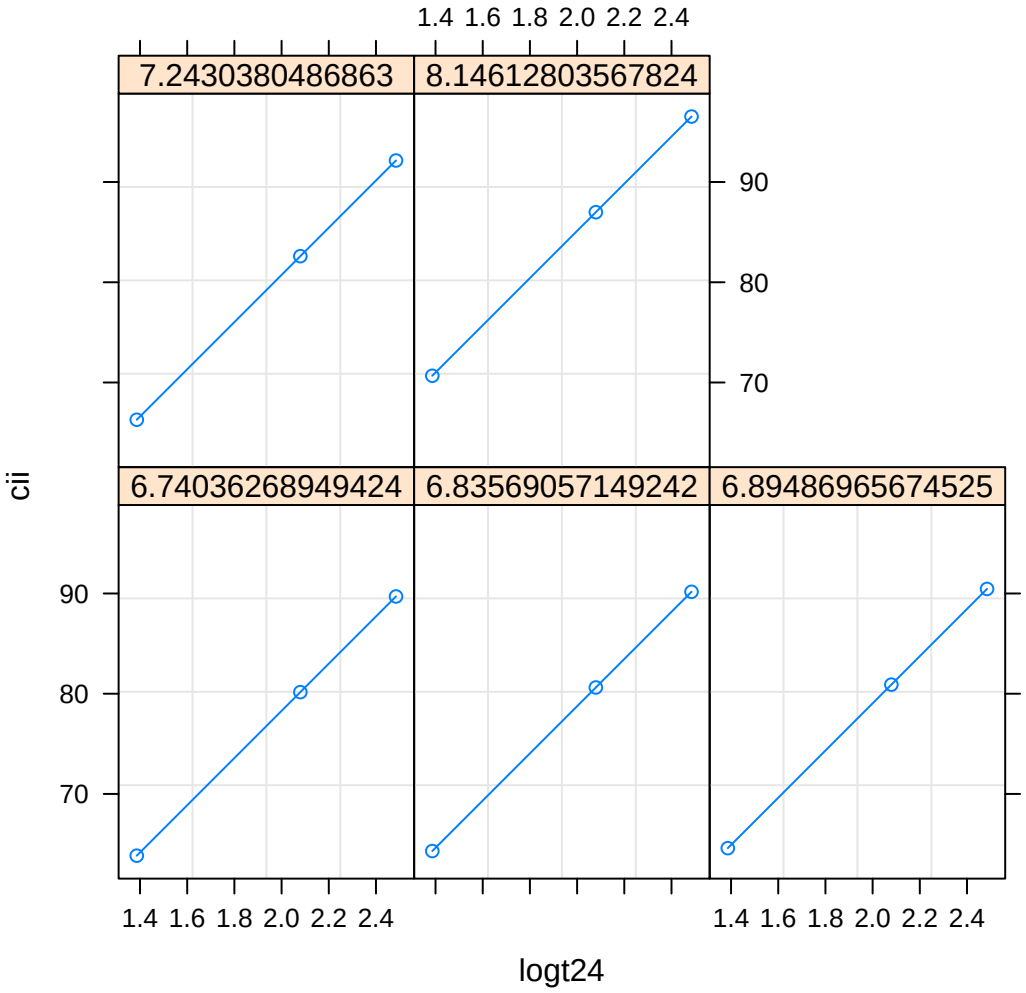
Beaucoup de travaux dans le domaine médical se sont apparus ces dernières années et dont la validité théorique des résultats établis a été basée sur l'interprétation

du coefficient de détermination. D'abord il faut noter deux choses : Le $\mathbf{R}^2$ n'est équivalent au ρ^2 que dans le cas de la regression linéaire simple, et les résultats obtenus peuvent seulement donner des indications, mais n'expriment pas des relations de causalités. Autrement dit une liaison n'entraine pas forcement une causalité.
Le $\mathbf{R}^2$ ne permet pas de dire si un modèle est bon ou pas. Une valeur faible de ce coefficient nous permet tout simplement de dire que le modèle établi est incomplet ou l'intensité du bruit est très élevée. Ainsi, comme le montre les résultats de l'ajustement du modèle sur les données la valeur du $\mathbf{R}^2_{ajuste}$ apprécie bien le modèle établi.

Pour tester si le modèle est bon ou pas, c'est en moment qu'intervient la statistique de **Fischer** notée **F**. Vue sa valeur, on peut décider laquelle des deux hypothèses est bonne.

$$\begin{cases} H_0 & \text{C depend uniquement de de i où t} \\ \text{vs} & \\ H_1 & \text{C depend des deux variables} \end{cases}$$

Au seuil de 0.05, on peut décider si l'on a réellement un bon modèle ou pas. Les résultats fournis par nos deux ajustements ne laissent pas penser que la concordance des caractères soit uniquement dépendante d'une des variables.
La capacité prédictive du modèle a été évaluée en appliquant la prédiction sur les données qui ont servi à l'ajustement du modèle, et si on définit l'erreur de prédiction comme étant la distance moyenne entre les valeurs ajustées et les valeurs observées, on trouve encore au seuil 0.05 une valeur de 0.0279. La figure suivante montre les concordances prédites entre 4 et 12 h par le modèle sur les données l'Escherchia coli. Les variables *cii* et $log24$ contiennent respectivement les concordances et les temps d'incubation correspondant à $4, 8$, et 12.



Le cas de Salmonelle paratyphi A

Le modèle de nouveau a été ajusté sur le jeu de données de l'espèce Salmonelle paratyphi A. Les paramètres estimés sont :

$\hat{\alpha} = 5.029$, $\hat{\beta} = 7.531$, $\mathbf{R}^2 = 0.77$, $\mathbf{R}^2_{ajuste} = 0.75$, et $\mathbf{F} = 2.7810^{-6}$

les tests de significativité associés aux paramètres sont :
α **P-value** $= 0.0027$
β **P-value** $= 2.6710^{-6}$
Sur ce jeu de données le test de significativité du paramètre α n'a pas nécessité la suppression de l'information obtenue au bout de 24 h.
L'erreur de prédiction sur ce nouveau ajustement est de 0.0313 ce qui est acceptable au seuil de 0.05

3.3 Examen des hypothèses sur le modèle d'erreur

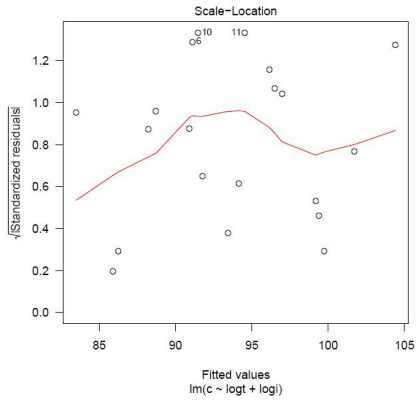
Lorsque les paramètres sont estimés et le modèle validé, il est nécessaire de s'assurer que les hypothèses sous-ajacentes à l'ajustement d'un modèle d'erreur additif, homoscédastique et gaussien, sont respectées. Cette réponse se fonde sur l'examen des résidus. Les résidus ($\hat{\varepsilon}_i$) sont les écarts entre valeurs observées et valeurs prédites par le modèle, ils sont égaux à la somme de l'erreur ε_i et du défaut de modélisation $(X\theta)_i - (X\hat{\theta})_i$.

$$\hat{\varepsilon}_i = C_i - \hat{C}_i = C_i - (X\hat{\theta})_i = (X\theta)_i + \varepsilon_i - (X\hat{\theta})_i = \varepsilon_i + (X\theta)_i - (X\hat{\theta})_i$$

Cette expression montre que les résidus ne sont pas indépendants car s'expriment tous en fonctions des paramètres. Une concordance $C_i = 80$ ajustée par 90 a un résidu de 10. Elle est cependant mieux représentée qu'une concordance $C_i = 5$ ajustée par 14 qui a un résidu de 9. Ces difficultés font que les résidus bruts ne sont pas interprétables (ils n'ont pas en général la même variance). Pour solutioner ce problème, on utilise des versions standardisées de ces résidus. Cependant dans la pratique il d'usage d'utiliser la version studentisée ou version réduite car il se trouve qu'en général, la version standardisée ne tient pas compte de la i^{eme} observation. Dans **R** on les obtient par la fonction *rstudent*. Toute analyse des résidus sera fondée sur les résidus réduits qui sous les hypothèses usuelles, doivent ressembler à une suite de variables aléatoires indépendantes, de même loi centrée et de variance 1.

Homoscédasticité des erreurs

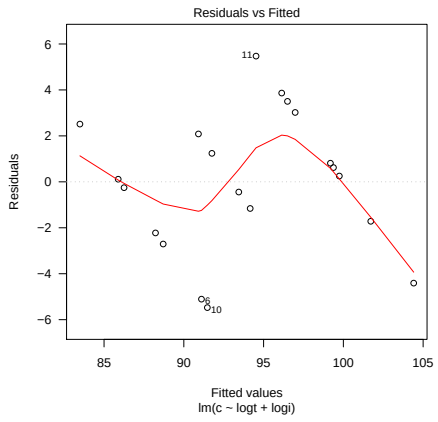
Il est nécessaire de tester si le modèle de variance est correct. Ceci est en général réalisée par un examen des résidus en fonction de la réponse prédite. Le cas des dénombrement bactéreins pose ne général un problème d'hétéroscédasticité car au cours d'une croissance, la densité de biomasse augmente fortement.



D'aprés ce grahique de contrôle, aucune variation évidente de l'amplitude des résidus n'ayant pas été observée au cours de la croissance.

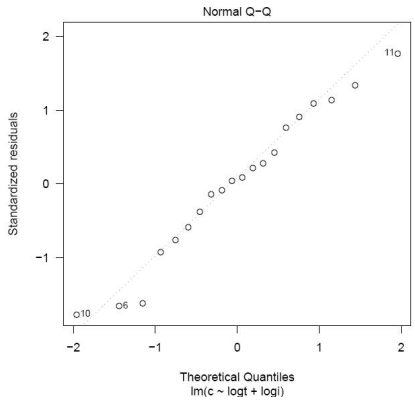
Résidus indépendants et centrés

Il est également possible d'utiliser le graghe les résidus réduits en fonction de la réponse prédite pour valider l'hypothèse d'un modèle d'erreur nulle. En effet une répartition équilibrée de part et d'autre de 0 en chaque zone du graphique peut être appréciée visuellement. Au cas où le graphique ferait apparaître une forme particulière, l'hypothèse d'indépendance pourrait être également mise en doute.



Normalité des erreurs

L'hypothèse de normalité peut être testée en utilisant la proximité entre les résidus réduits et une suite de variables aléatoire gausienne (indépendantes, centrées, réduites). Un tel diagnostic est facilité par l'emploi d'un diagramme quantile-quantile (qqplot) : représentation des quantiles de la distribution empirique des résidus en fonction des quantiles exacts de la loi normales centrée réduite. Le nuage de points ainsi obtenu est réparti selon une droite, l'hypothèse de normalité des résidus ne peut pas être rejetée.



3.4 Une analyse globale de l'identification des espèces par le temps d'incubation

3.4.1 Première approche

L'exemple des données de l'Escherichia coli

Avec l'analyse faite sur les données de l'E.coli, par le modèle logit, on avait dit que la variable inoculum ne sembler contenir aucune information pour la variable *MAN*. En effet on a :
La réponse de la variable *MAN* 5 fois les valeurs $0, 1, 1, 1$
La variable explicative t 5 fois les valeurs $4, 8, 12, 24$
La variable explicative inoculum :
$*$ 4 fois la valeur 5.510^6
$*$ 4 fois la valeur 6.8510^6

$*4$ fois la valeur 7.6510^6
$*4$ fois la valeur 1.7510^6
$*4$ fois la valeur 1.410^6

Pour toute valeur de l'inoculum considérée, le test *MAN* est négatif si $t = 4$ et est positif si $t = 8, 12, 24$. Ainsi la variable explicative t semble contenir l'essentiel de l'information pour la variable réponse *MAN*.

Cependant, en faisant une analyse globale des différents tests biochimiques, on peut constater qu'il y'a nul besion d'un quelconque modèle log-linéaire pour prédire les caractères biochimiques par les variables t et *inoculum* : le problème est déterministe.
On peut constater que les tests *ADH*,*CC*,*CS*,*GEL*,*H2S*,*MAL*,*LDC*,*URE* sont tous négatifs durant l'incubation et pour toute valeur de l'inoculum considérée.
Les tests *ONPG* et *ADH* sont négatifs si $t = 24$.
Le test *SOR* est positif si $t = 24$ ou $t = 12$ et l'inoculum dépassant la valeur 6.8510^6.
Les tests *VP* et *PDA* sont négatifs si $t = 24$.
Le test *GLU* est négatif si $t = 4$ et l'inoculum atteignant 5.510^6 ou 6.8510^6.
Le test *LAC* est négatif si $t = 4$ et un inoculum strictement plus petit que 1.410^8.
Le test *XYL* est négatif si $t = 4$ ou $t = 8$ et l'inoculum strictement plus petit que 6.8510^6.
La variable explicative t est le principale facteur et plus particulièrement les valeurs extrèmes $t = 4$ et $t = 24$.

On peut proposer aux biologistes d'éliminer les caractères qui ne varient pas durant l'incubation (*ADH*,*CC*,*CS*,*GEL*,*H2S*,*MAL*,*LDC*,*URE*) et de s'intéresser à :
L'équivalence des tests *ONPG*, *VP*,*IND* et *PDA*.
L'équivalence des tests *ODC* et *MAN*.
L'évolution des caractères *GLU*,*LAC*,*SOR* et *XYL*.

3.4.2 Deuxième approche

Dans cette partie on suppose que par des études antérieures on a obtenu des mesures sur les quatre espèces dont porte l'étude. Dans la suite on cherche à faire une discrimination entre E.coli, pneumonia, Salmonella et Acitenobacter. Pour E.coli, pneumonia et salmonella on a le même nombre et les mêmes tests, tandis que pour Aciteno les tests *VP* et *LDC* ont été respectivement remplacés par le *ESC* et le *NIT*, et on note l'absence du test *ODC* .
Pour pouvoir comparer les données, on a ajouté les caractères manquants à certain tableaux de données. Ainsi pour Aciteno on a ajouté les tests *VP*,*LDC* et *ODC* avec NA (Not Available : valeur manquante), et tous les caractères indetérminés dans les bases de données ont été également substitués par NA, considérant que ce sont des valeurs manquantes. Le *NIT* a été lui aussi ajouté aux données de E.coli, pneumonia et Salmonella, soit au total 19 caractères dans l'ordre :*ADH*,*ONPG*,*CC*,*CS*,*VP*,*NIT*,*GEL*, *H2S*,*IND*,*MAL*,*PDA*, *LDC*,*ODC*, *URE*,
GLU,*LAC*,*MAN*,*SOR*,*XYL*.
IL y'a donc $19 * 4 * 5 = 380$ mesures pour chacune des 4 espèces dans l'ordre des temps d'incubation et des inocula. On a les mêmes temps d'incubation, mais tel

n'est pas le cas pour les inocula et la différence entre ces derniers pour certaines espèces est significative (voir figure : comparaison des densités). Cependant, dans une première approximation, on ne tient pas compte des tailles des inocula. Comme on l'a évoqué plus haut, avec les données que nous avons à faire, le problème n'est pas statistique mais plutôt déterministe, et on peut le voir de façons :

Première méthode

On réduit les données aux résultats de la 4^e heures d'incubation : on enléve aux vecteurs (variables contenant les résultats des 380 mesures de chacune des 4 espèces dans l'ordre des temps d'incubation et des tailles des inocula) de longueur 380 les résultats des heures $24, 12, 8$: il reste des vecteurs de longueur $95 = 19 * 5$. On compare ces vecteurs 2 à 2 et on enléve les résultats NA ou d'égalités . Dans le programme **R** les résutats d'égalités correspondent aux FALSE. Aprés nettoyage des données (suppression des NA et des résultats d'égalités), il reste 5 caractères :*GLU*,*LAC*,*MAN*,*SOR* et *XYL*.

Pour Aciteno et E.coli, les tests *MAN*,*SOR* et *XYL* sont négatifs pour toutes les mesures ; pour pneumonia et Salmonella, ces derniers sont tous positifs (exepté un seul). Pour Aciteno, le test du glucose (*GLU*), est négatif pour toutes les mesures tandis qu'il est positif pour les 3 mesures d'inocula les plus élevés de E.coli. Notons au passage que les valeurs des inocula sont très comparables pour ces deux dernières espèces.
Pour pneumonia, le test du glucose et du lactose (*LAC*) sont respectivement positif et négatif pour toutes les mesures, pour Salmonalla le test du glucose est négatif pour toutes les mesures sauf la dernière et le test du lactose est positif pour toutes les mesures.

On peut donc parfaitement identifier les 4 espèces dès le temps d'incubation $t = 4h$ à l'aide de ces 5 caractères avec des inocula de taille 10^7 et cela avec redondance (pour comprendre cette partie il nécessaire d'exécuter le programme **R**).

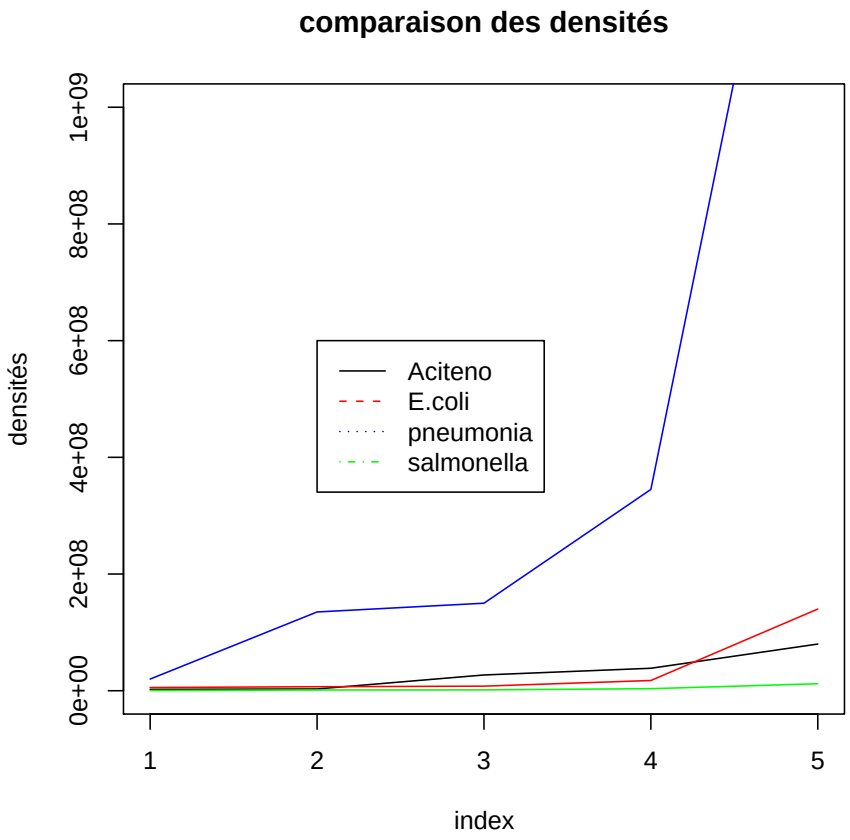
Une partie des résultats de l'exécution du programmme **R** correspondant .

```
ac4=(acito4!=coli_4)
ap4=(acito4!=pmoni4)
as4=(acito4!=salmo4)
cp4=(coli_4!=pmoni4)
cs4=(coli_4!=salmo4)
ps4=(pmoni4!=salmo4)

Resultats (après nettoyage des NA et suppression des égalités (FALSE)

names(ac4)[ac4]
"glu" "glu" "glu"
"lac"

names(ap4)[ap4]
"glu" "glu" "glu" "glu" "glu"
```

```
"man" "man" "man" "man" "man"
"sor" "sor" "sor" "sor" "sor"
"xyl" "xyl" "xyl" "xyl" "xyl"

names(as4)[as4]
"glu"
"lac" "lac" "lac" "lac" "lac"
"man" "man" "man" "man" "man"
"sor" "sor" "sor" "sor" "sor"
"xyl" "xyl" "xyl" "xyl"

names(cp4)[cp4]
"glu" "glu"
"lac"
"man" "man" "man" "man" "man"
"sor" "sor" "sor" "sor" "sor"
"xyl" "xyl" "xyl" "xyl" "xyl"

names(cs4)[cs4]
"glu" "glu"
"lac" "lac" "lac" "lac"
"man" "man" "man" "man" "man"
"sor" "sor" "sor" "sor" "sor"
"xyl" "xyl" "xyl" "xyl"

names(ps4)[ps4]
"glu" "glu" "glu" "glu"
"lac" "lac" "lac" "lac" "lac"
"xyl"

acito4
  glu  glu  glu  glu  glu
    0    0    0    0    0
  lac  lac  lac  lac  lac
    0    0    0    0    0
  man  man  man  man  man
    0    0    0    0    0
  sor  sor  sor  sor  sor
    0    0    0    0    0
  xyl  xyl  xyl  xyl  xyl
    0    0    0    0    0

coli_4
  glu  glu  glu  glu  glu
    0    0    1    1    1
  lac  lac  lac  lac  lac
```

```
    0    0    0    0    1
  man  man  man  man  man
    0    0    0    0    0
  sor  sor  sor  sor  sor
    0    0    0    0    0
  xyl  xyl  xyl  xyl  xyl
    0    0    0    0    0

pmoni4
  glu  glu  glu  glu  glu
    1    1    1    1    1
  lac  lac  lac  lac  lac
    0    0    0    0    0
   man  man  man  man  man
     1    1    1    1    1
   sor  sor  sor  sor  sor
     1    1    1    1    1
   xyl  xyl  xyl  xyl  xyl
     1    1    1    1    1

salmo4
  glu  glu  glu  glu  glu
    0    0    0    0    1
  lac  lac  lac  lac  lac
    1    1    1    1    1
  man  man  man  man  man
    1    1    1    1    1
  sor  sor  sor  sor  sor
    1    1    1    1    1
  xyl  xyl  xyl  xyl  xyl
    1    0    1    1    1


Conclusions à ti=4
------------------
1)
  man = 0, sor = 0, xyl = 0 ==> acito ou coli

  glu = 0                   ==> acito
  glu = 1 (i assez grand)   ==> coli

2)
  man = 1, sor = 1, xyl = 1 ==> pmoni ou salmo

  glu = 1, lac = 0          ==> pnomi
  glu = 0, lac = 1          ==> salmo
```

Les variables acito4, coli4, pnomi4 et salmo4 contiennent les résultats des mesures reduites à la $4^e h$ d'incubation. Les variables ac4, ap4, as4, cp4, cs4, ps4 permettent de faire la comparaison des vecteurs 2 à 2. La fonction **names(ca4)**[$ac4$] par exemple permet de récupérer les tests qui sont différents (elle récupére les nons correspondant aux valeurs logique TRUE dans le programme).



Deuxième méthode

Cette méthode est un résultat de l'observation des données.
Si l'on supprime les résultats au temps $t = 24h$, les 9 variables *ADH*,*ONPG*,*VP*,*NIT*,*GEL*, *H2S*, *IND*,*MAL*,*PDA* sont égales sur les 4 espèces.
Si l'on supprime en outre les résultats au temps $t = 12h$, les 2 variables *URE*,*LDC* sont égales sur les 4 espèces.
Si l'on supprime enfin les données du temps $t = 8h$, les 3 variables *CC*,*CS*,*ODC* sont égales sur les 4 espèces, et on retrouve le résultat précédent : il reste les 5 variables *GLU*,*LAC*,*SOR*,*MAN*,*XYL* qui discriminent de façon exacte les 4 espèces.

Discussion

Notre travail a porté sur les données des souches qui ont été identifiées et conservées par l'équipe du laboratoire bactériologie de l'Hopital Aristide le Dantec.

En microbiologie prévisionnelle, la démarche suivie pour décrire l'évolution d'une population bactérienne est la construction d'un modèle primaire. Une altérnative aux techniques classiques de modélisation pour la prédiction de la croissance microbienne à partir des premiers points expérimentaux d'une cinétique a été proposée. En pratique, chaque problème de prédiction de la croissance microbienne présente ses propres difficultés et exigences. La technique d'inférence bayésienne proposée permet d'ajuster des mesures précoces d'une cinétique, et elle intéressante dès lors qu'un quelconque paramètre est connu a priori de façon floue ou en présence de plusieurs données manquantes.

Notre étude, loin être exhaustive, correspond ainsi à la proposition de pistes de recherches. D'aures méthodes mériteraient certainement d'être tester comme les techniques de réseaux de neurones artificiels, les réseaux de neurones récurrents . Cette étude met en évidence également une des capacités intéressante de la théorie du JACKKNIFE qui est la facilité qu'elle confère à la manipulation de fonctions des paramètres. En effet, la taille des jeux de données n'exclue pas la présence d'un biais potentiel dans l'estimation des paramètres des modèles proposés. Elle empêche également de profiter des avantages tels que l'utilisation des modèles généralisés de régression linéaires ou non linéaire. La théorie du Jackknife peut être utilisée non seulement dans l'estimation des paramètres, mais également peut permettre une estimation par intervalle de ces différentes fonctions de paramètres. Si seuls les résultats de la théorie de la régression non linéaire sont considérés, l'estimation par intervalle d'une fonction des paramètres ne peut être obtenue qu'au prix de la linéarisation de la partie deterministe du modèle de régression et de la connaissance de la loi du vecteur des paramètres. L'identification de la loi d'un paramètre n'étant pas évidente sauf dans le cas où la taille du jeu de données est très grande, la loi du veteur des paramètres est approchée asymptotiquement. Dans le cas de petits jeux de données, le jackknife, sans faire d'hypothèses de loi, permet l'estimation par intervalle d'une fonction quelconque des paramètres. L'utilisation du jackknife permet aussi d'évaluer la robustesse des modèles établis : il s'agit d'évaluer la sensibilité du modèle vis à vis de variations de jeux de données. Le jackknife apparaît donc comme un outil particulièrement adapté pour les petites bases de données. D'ailleurs, lorsque les échantillons sont de grande taille les résultats obtenus selon la théorie du jackknife convergent vers ceux obtenus par la théorie de la régression non linéaire.

La méthode proposée dans cette étude pour modéliser la concordance des carac-

tères biochimiques en fonction des inocula et du temps d'incubation a nécessité :
− L'utilation d'un modèle logit qui n'a pas pu adapter aux jeux de données. Les résultats obtenus par son analyse semblent prometteurs, cependant il est nécessaire de considérer une serie de mesures raffinées dans l'ensemble $\{4, 5, \ldots, 24\}$ ce qui constitue une grande variabilité biologique difficilement contrôlable par les manipulateurs.
− La construction d'un modèle de régression linéaire multiple avec une bonne capacité prédictive au seuil de 0.05. Malgré le petit nombre de points de mesures, l'hypothèse d'un modèle d'erreur additif homoscédastique n'a pas pu être rejetée .
− Une analyse par cas des espèces qui a obouti essentiellement au besion pour les biologistes de s'intéresser à l'évolution des caractères *GLU*,*LAC*,*SOR*, *XYL*, d'étudier l'équivalence de *ONPG*, *VP*,*IND*, *PDA*, l'équivalence *ODC* et *MAN*.
− Une analyse globale des mesures des espèces pour des fins discriminatoires. Les différents espèces ont pu être identifier à $t = 4$ h avec un inoculum d'environ 10^7.

Dans le cadre de la modélisation appliquée au domaine des sciences du vivant, il semble important de considérer l'existence potentielle d' erreurs de mesures et de facteurs de confusions. Dans le cas où les différents erreurs de mesures et facteurs expérimentaux confondus sont identifiés, la théorie des plans d'expériences permet de gérer ce problème. Cependant dans le domaine des sciences du vivant, il semble difficile d'identifier tous ces facteurs. La confusion de facteurs et les erreurs de mesures peuvent avoir une incidence dans le cadre de l'utilisation d'un modèle mathématique à des fins prévisionnelles. Ainsi on voit toute l'importance du dialogue, malheureusement absent le plus souvent entre les deux acteurs de la modélisation appliquée à la microbiologie : le microbilogiste et le modélisateur.

Les résultats déduit de cette étude ne sont valables que dans le cadre du travail exposé dans ce rapport. Natamment le fait les profils des évolutions biochimiques des différents espèces sont fonction du temps d'incubation et de la taille des inocula.

Conclusion

La problématique de l'identification de pathogènes au laboratoires ne constitue plus un problème pour les microbiologistes. En effet divers galeries d'identification d'une très grande rapidité et efficacité sont parties en concurrence sur le marché. Cependent une frange importante de la population des pays en voie de développement, qui pourtant est la plus exposée aux infections causées par ces germes n'accéde pas encore à ces galeries du fait de leur coût élevé. C'est dans ce contexte que le laboratoire bactériologie -virologie fondementale et appliquée à initier les micro-plaques CSB, de moindre coût et d'efficacité garantie pour l'identification des espèces. L'amélioration des maux dont souffrent encore ces micro-plaques : le manque de la standardisation de l'inoculum et le temps d'incubation long de 24 h ont été les objectifs majeurs développés dans ce travail.

Dans le but d'une contribution à l'amélioration de ces micro-plaques CSB, une méthode intégrant une information biologique fiable a été proposée pour prédire tout le reste de la cinétique à partir de ses premiers points expérimentaux par l'étude de la croissance. On a montré sur des exemples ponctuels qu'elle peut donner des résultats intéressants. Différentes méthodes d'analyses ont été proposées aux biologistes pour des fins discriminatoires des espèces dès le temps le d'incubation $t = 4$ h avec un inoculum d'environ 10^7 UFC.

Néamoins l'étude est lion d'être exhaustive car on n'a pas pu proposer une méthode permettant à chaque inoculum fixé de donner le profil biochimique de l'espèce pour tout temps d'incubation. Malheureusement cette technique ne peut voir le jour que si la serie de mesures est raffinée. Les techniques de réseaux de neurones en plein expansion dans l'industrie peuvent être comparer à la technique d'inférence bayésienne utilisée dans ce cadre précis. Une bonne technique de dénombrement des cellules viables en mettant en exergue les erreurs de mesures et éventuels facteurs confondants pourrait certainemment donner une excellente analyse du temps d'incubation par l'utilisation d'un réseaux de neurones recurrents.

A l'heure actuelle où on prône une médecine basée sur des preuves démontrées, l'utilisation des mathématique et l'informatique est amenée à se développer. Les mathématiciens et informaticiens sont appelés à mettre leurs compétences au service de la médecine et de la biologie. Cependant, ils doivent s'efforcer de comprendre leurs besoins afin de pouvoir développer des méthodes permettant d'y répondre.

ABREVIATIONS

UCAD : Univérsité Cheikh Anta Diop
API : Appareil et procédé d'identification
ADH : Arginine dihydrolase
ADN : Acide desoxyribonucléique
ADO : Adonitol
ADP : Adénosine dilphosphate
ARN : Acide ribonucléique
CC : Citrate de christensen
CS : Citrate de Simmons
ESC : ESC
GEL : Gélatine
GLU : Glucose
H2S : Sulfure d'hydrogène
IND : Indole
LAC : Lactose
LDC : Lysine décarboxylase
MAL : Malonate
MAN : Mannitol
NIT : Nitrate
ODC : Ornithine décarboxylase
ONPG : Orthonitrophényl
PDA : Phénylalanine désaminase
SAC : Saccharose
SOR : Sorbitol
VP : Voges proskaauer
XYL : Xylosus
ie : c'est à dire
MLE : Maximun likehood method (méthode du maximun de vraissemblance)
A' : Transposée de A

BIBLIOGRAPHIE

[1] S. M.Fatou utilisation des méthodes biométriques pour la validation de l'identification des occi à Gram positif, *thèse pharm.,Dakar,*2007.

[2] G. Oulimatou utilisation des méthodes biométriques pour la validation de l'identification des occi à Gram négatif, *thèse pharm.,Dakar,*2007.

[3] M. M.Erneville étude de l'effet de l'inoculum et du temps d'incubation sur l'identification des bacteries à Gram négatif, *thèse pharm.,Dakar,*2008.

[4] S. Breand étude biométrique de la réponse d'une population bactérienne à une variation défavorable de température ou de PH.Application en microbiologie prévisionnelle alimentaire,labotatoire biométrie-génétique et biologique des population UMR CNRSS 5558, *thèse université Claude Bernard Lyon I,*1998 .

[5] M. Cornu dynamique des population bactérienne en culture mixtes,laboratoire biométrie et biologie évolutive UMR CNRSS 5558, *thèse université Claude Bernard Lyon I,*2000 .

[6] M. L.D.Muller méthodes de prédictions des aptitudes de croissance des populations de micro-organismes ,laboratoire de biométrie-génétique et biologie des populations, *thèse université Claude Bernard Lyon I,*1995 .

[7] P. Lezand les modèles linéaires à effet mixtes en pratique :Analyse de la perception du risque de conflit par un contrôleur aérien.Equipe de recheche en mathématiques appliquées (SDER/ENAC),séminaire LAPMA 27 mai 2005

[8] Method validation in comformity with standard ISO/CEI 17025,laboratoire natinal d'essais.

[9] M.Cornu,M.L.D.Muller,J. P.Flandrois characterization of unexpected growth of Escherchia coli O 157 :H7 by modeling,appplied and enviromental microbiology,Dec 1999,P.5322 − 2327.

[10] M. Sanaa microbiologie prévisionnelle :principaux modèles de croissance utiles en appréciation quantitatives des risques ,école national vétérinaire d'Alfort,france

[11 L. Vézina,M. Lacroix système Biolog :identification des bacteries aéribies Gram négatifs,laboratoire de diagnostic en phytoprotection,Québec

[12] J. Schindler,Z. Schindler numerical identification of bacteria with a hand -held calculator as an alternative to code books ,journal of clinical microbiologie ,Feb 1982,p.$332 - 334$

[13] Langham,C.D,Sneath,P.H.A,Williams,S.T.Mortimer,A.M1989 detecting aberrant strains in bacterial groups as an aid to constructing databases for camputer identification,journal of applied bacteriology 66,$339 - 352$.

[13] B. Van Oystaeyen aide numérique à l'identification bactérienne : définition d'un risque d'erreur alpha.,ann biol clin 2006,64(1) :$83 - 89$.

[14] M. Gyllenberg,T. Kosski probabilistic models for bacterial toxonomy,international statistical review (2001),69,2,$249 - 276$,printed in Mexico by international statistical institute.

[15] J.C. Thalabard cours épidémiologie Master2 STAFAV université Gaston Berger de Saint-Louis du Sénégal,2008.

[16] P. Ngom cours MCMC and statistical decision thoery and bayesian analysis Master2 STAFAV université Gaston Berger de Saint-Louis du Sénégal,2008.

[17] J. Coursol cours Data mining et base de données Master2 STAFAV université Gaston Berger de Saint-Louis du Sénégal,2008.

[18] J. J.Drosesbeke,M. Lejeune,G. Saporta modèles statistiques pour données qualitatives,2005,*Edition Technip*

[19] F. Rossi cours réseaux de neuronnes :modèles linéaires généralisé,université paris-IX Dauphine.

[20] J.C. Bertrand écologie bactérienne marine,cours DEA biosciences et environnement,chimie et santé.

[21] G. Pina,D. Raynaud critéres de choix d'une méthode d'identification cours DES bactériologie-virologie, Grenoble,2003.

[22] G. Blanchette quelque généralisation du modèle de regression logistique,mémoire de maîtrise és sciences,université laval,Québec mai 1996.

[23] P. Taffe cours de regression logistique appliqué,institut universitaire de médecine sociale et préventive(IUMSP) et centre d'épidémiologie clinique (CEPIC),Lausanne,Août 2004.

[24] F. Duymé cours de regression logistique binaire institut supérieur d'agriculture de Lille.

[25] Health protection agency identification of Streptococcus species,enterococcus species and morphologicaly similar organisms,issued by standard unit,evaluation and standard laboratory.Specialist and reference microbiology division.

[26] Health protection agency identification of glycose non-fermenting rods,issued by standard unit,evaluation and standard laboratory.Specialist and reference microbiology division.

[27] Health protection agency identification of Staphylococcus species,micrococcus species and stomatococcus species,issued by standard unit,evaluation and standard laboratory.Specialist and reference microbiology division.

[28] C.L. Brosnikoff,R.P. Rennie,L.C. Turnbull evaluation of organism recovery,after twlve months using the roll plate method described in CLSI document M40 :quality control of microbiollogical transport devices,Medical microbiology laboratory,university of Alberta hospital,Edmond AB,canada.

[29] Technique de laboratoire pour le diagnostic des méningtes à neisseria meningitidis,Streptococcus pneumonia et haemophilus influenza,organisation mondiale de la santé,département maladies transmissibles,surveillances et action.

[30] T. Kajalainen,E. Ruokoboski,P. ukkonen,E. herve survival of Streptococcus pneumoniae,haemophilus,influenzae,and moraxella cathlis.Frozen in skin milk.Tryptone-Glucose,glycerol medium,journal of microbiology ,jan 2004,p.412 − 414