Modèles à effets mixtes

Université de Lille

Master MISO

C. Dumont

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

"All models are wrong but some are useful" (G. Box).

Qu'est-ce qu'un modèle ?

- On cherche à trouver les lois qui gouvernent un processus donné.
- Un modèle est une structure générale définie par une fonction mathématique.
- Les données ne sont pas exactement égales aux prédictions du modèle en raison :
  - des approximations faites par le modèle ;
  - des erreurs de mesure.

  "All models are wrong but some are useful" (G. Box).

Données longitudinales (1/2)

- Données longitudinales : études où la réponse est observée pour chaque participant/groupe de façon répétée dans le temps.
- Les données longitudinales ont des caractéristiques différentes qui doivent être prises en compte.
- Des modèles et des méthodes adaptées à ces particularités sont nécessaires.

Données longitudinales (2/2)

- Concentrations au cours du temps (pharmacocinétique)
- Essais cliniques avec visites répétées
  - modélisation de biomarqueurs
  - évolution sous traitement
- Agronomie
  - modèles de croissance
- Génétique animale

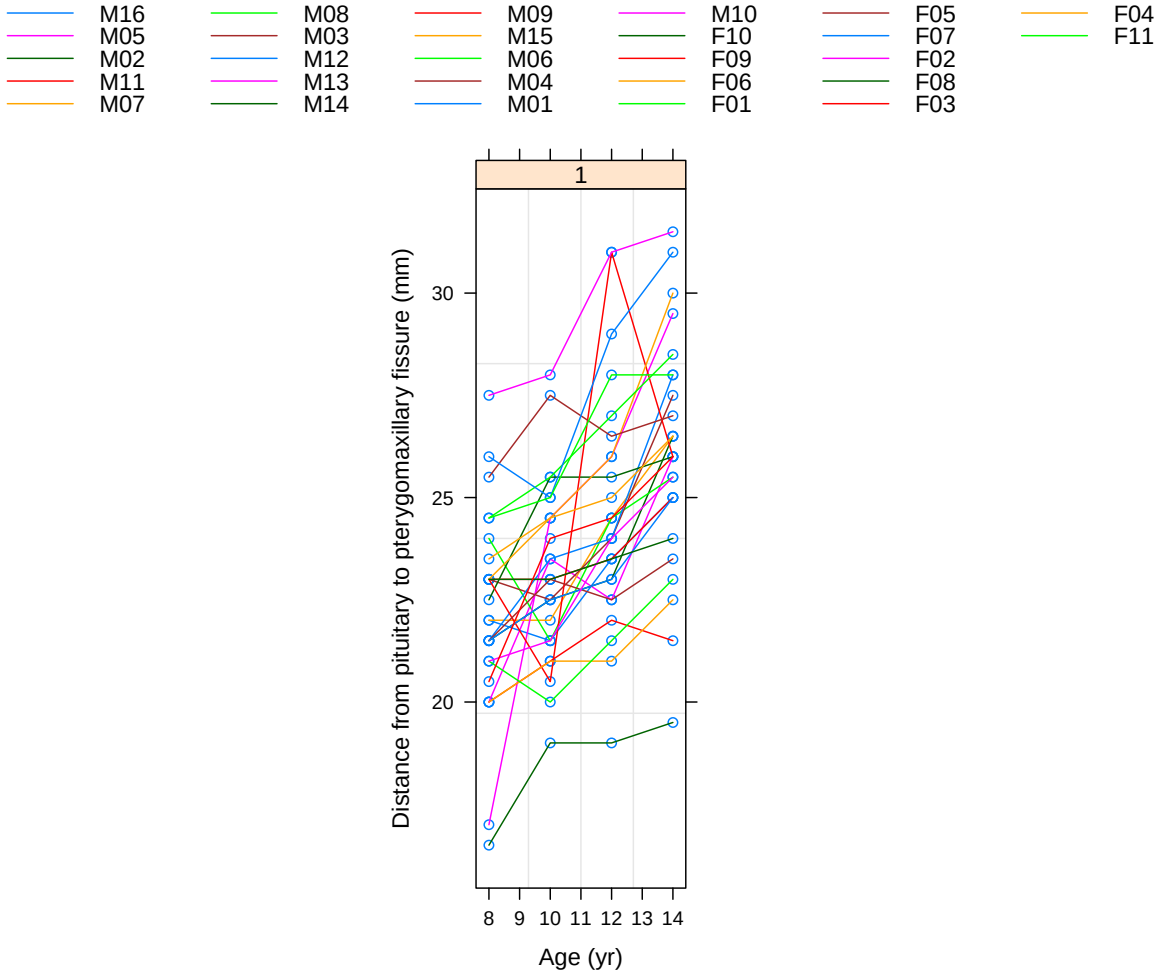
Exemple introductif : étude dentaire (Pothoff et Roy 1964) (1/6)

- Présentation de l'étude :
  - 27 enfants (16 garçons et 11 filles)
  - Pour chaque enfant, mesure de distance entre deux dents à 8, 10, 12 et 14 ans

- Questions :
  - Est-ce que la distance change avec l'âge ?
  - Quel est le modèle du changement ?
  - Est-ce que ce modèle diffère selon le sexe ? Et si oui, comment ?

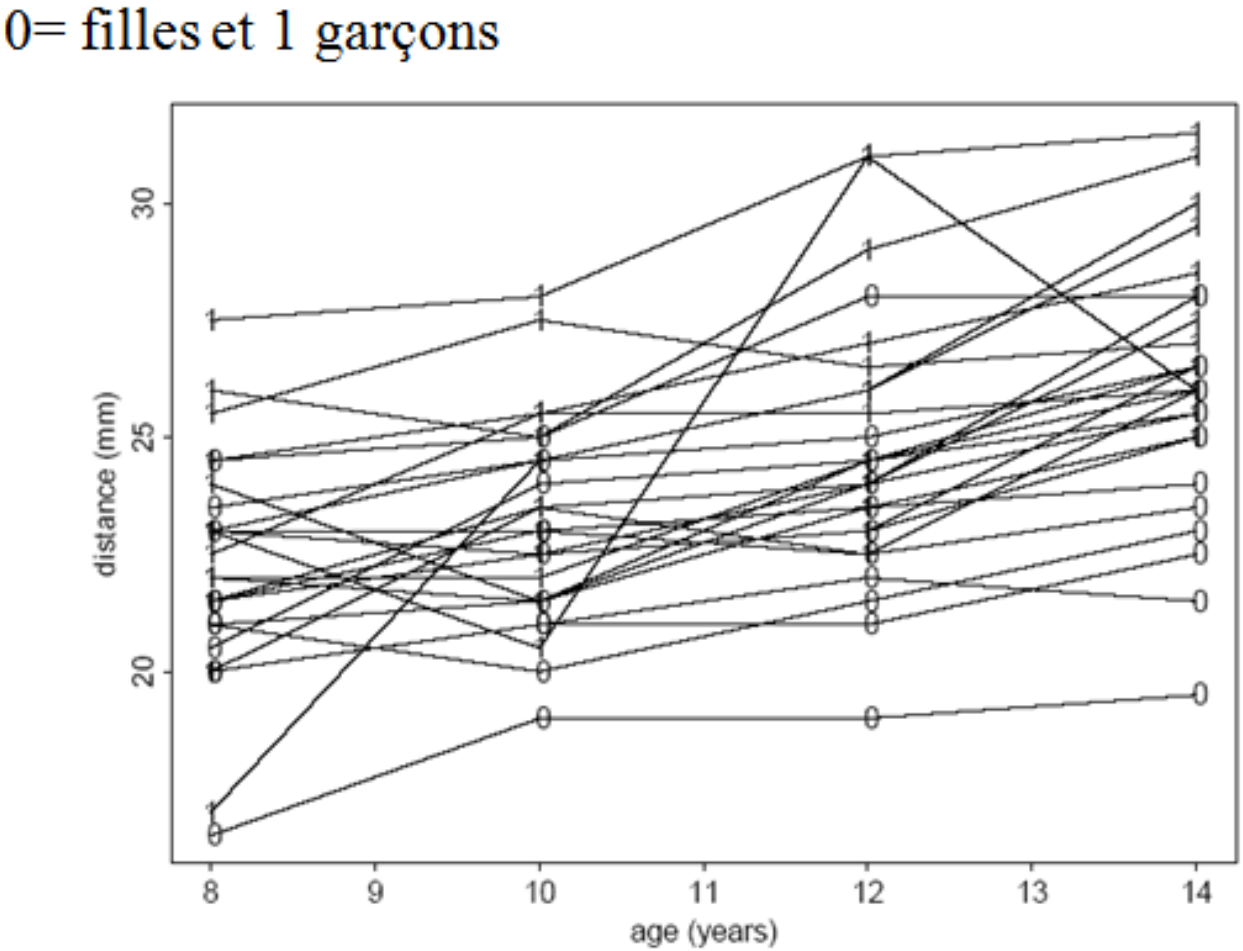
Exemple introductif : étude dentaire (Pothoff et Roy 1964) (2/6)

Données individuelles



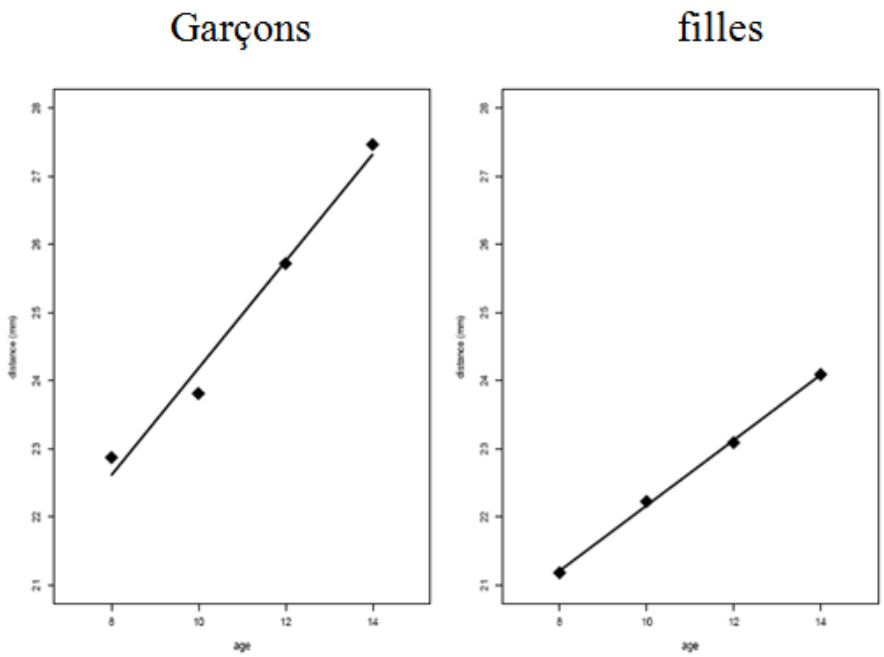
Exemple introductif : étude dentaire (Pothoff et Roy 1964) (3/6)

Données individuelles



Exemple introductif : étude dentaire (Pothoff et Roy 1964) (4/6)

Distances moyennes à chaque âge par sexe



Remarques :

- Tous les enfants ont des mesures aux mêmes dates.
- Les modèles par enfant (individuels) ressemblent au modèle moyen (ligne droite croissante).

Exemple introductif : étude dentaire (Pothoff et Roy 1964) (5/6)

Approche "Naive Pooled Data" puis régression linéaire classique

Deux questions qu'on peut se poser :

- Est-ce que la pente de la courbe est la même selon le sexe ?
- Comment évolue la moyenne au cours du temps ?

Pour tous les sujets i à l'âge t_{ij}, on a le modèle linéaire suivant :
$y_{ij} = \beta_{0F} + \beta_{1F} \times t_{ij} + \varepsilon_{ij}$, si i fille
$y_{ij} = \beta_{0G} + \beta_{1G} \times t_{ij} + \varepsilon_{ij}$ sinon.

- Ajuster le modèle sur l'ensemble des données (en oubliant les provenances individuelles) avec la méthode des moindres carrés ordinaires et tester si $\beta_{1G} = \beta_{1F}$.
- Problème car cette méthode ignore les corrélations au sein d'un même sujet donc biais lorsqu'on utilise la méthode des moindres carrés ordinaires (ou le maximum de vraisemblance).

Exemple introductif : étude dentaire (Pothoff et Roy 1964) (6/6)

Chaque enfant a sa propre trajectoire (rectiligne).
Est-ce que le modèle est différent selon les garçons et les filles ?

- Est-ce que la trajectoire "typique" (moyenne) est la même indépendamment du sexe ?
- On pourrait estimer les paramètres pour chaque sujet puis faire un résumé statistique et des tests :
  - problème des erreurs d'estimation (négligées)
  - nécessité d'avoir des protocoles assez équilibrés entre patients
  - fastidieux
- Une analyse conjointe de tous les profils à travers le temps
  - Chaque sujet fluctue autour d'une tendance "moyenne"
  - Des erreurs de mesure peuvent se produire

Modèle à effets mixtes (1/4)

Le modèle à effets mixtes :

- modélise le comportement individuel ;
- permet d'analyser simultanément les données recueillies chez l'ensemble des sujets et notamment des données individuelles éparses ;
- permet d'estimer les paramètres moyens, leur variabilité inter-individuelle et l'effet d'éventuelles covariables.

Modèle à effets mixtes (2/4)

Pour un individu i, l'observation mesurée au temps t_{ij}, notée y_{ij} est définie par :

$$y_{ij} = f(\theta_i, t_{ij}, z_i) + g(\theta_i, t_{ij}, z_i)\varepsilon_{ij}$$

On considère N sujets $i = 1, ..., N$ disposant de n_i observations.

On note $y_i = (y_{i1}, ...y_{in_i})$ le vecteur des n_i observations du sujet i.

θ_i : paramètres individuels du modèle pour le sujet i (vecteur de dimension p).
t_{ij} : temps correspondant à la $j$ème observation pour le sujet i.
z_i : vecteur des covariables (poids, âge, sexe, créatinine,...) qui peuvent varier au cours du temps.
ε_{ij} : erreurs résiduelles supposées indépendantes et identiquement distribuées selon une distribution gaussienne de moyenne nulle et de variance 1, *i.e.* $\varepsilon_{ij} \sim N(0, 1)$.

Modèle à effets mixtes (3/4)

f et g sont des fonctions des paramètres individuels.

f correspond au modèle structurel (linéaire ou non linéaire) choisi pour représenter le processus.

g correspond à la forme du modèle d'erreur, qui est supposée connue.
Un modèle fréquemment utilisé est le modèle suivant (dit modèle combiné) :

$$g(\theta_i, t_{ij}, z_i) = a + bf^c(\theta_i, t_{ij}, z_i),$$

qui comprend deux situations fréquentes :

- Lorsque $b = 0$ et $c = 1$, erreur constante ou homoscédastique de variance $g^2 = a^2$.
- Lorsque $a = 0$ et $c = 1$, erreur proportionnelle aux prédictions du modèle, de coefficient de variation b.

Modèle à effets mixtes (4/4)

Les paramètres individuels θ_i peuvent être décomposés en des effets fixes β et des effets aléatoires individuels b_i.

Il est souvent supposé :

- un modèle additif pour chaque composante $\theta_{i(k)}$ de θ_i (distribution normale pour $\theta_{i(k)}$) : $\theta_{i(k)} = \beta_{(k)} + b_{i(k)}$, soit $\theta_{i(k)} \sim N(\beta_{(k)}, \omega^2_{(k)})$;
- ou un modèle proportionnel (distribution log-normale pour $\theta_{i(k)}$) : $\theta_{i(k)} = \beta_{(k)} \times exp(b_{i(k)})$, soit $ln(\theta_{i(k)}) \sim N(ln(\beta_{(k)}), \omega^2_{(k)})$, où $\omega^2_{(k)}$ représente la variance du $k$ème effet aléatoire.

On note d la dimension de b et $b = (b_{i(1)}, b_{i(2)}, ..., b_{i(d)})$ le vecteur des effets aléatoires chez le sujet i.
Les effets aléatoires b_i sont supposés suivre une loi multinormale : $(b_{i(1)}, b_{i(2)}, ..., b_{i(d)}) \sim N(0, \Omega)$ où Ω est la matrice de variance-covariance $d \times d$ des paramètres aléatoires. On suppose souvent Ω diagonale.

Modèle à effets mixtes (4/4)

Les paramètres individuels θ_i peuvent être décomposés en des effets fixes β et des effets aléatoires individuels b_i.

Il est souvent supposé :

- un modèle additif pour chaque composante $\theta_{i(k)}$ de θ_i (distribution normale pour $\theta_{i(k)}$) : $\theta_{i(k)} = \beta_{(k)} + b_{i(k)}$, soit $\theta_{i(k)} \sim N(\beta_{(k)}, \omega^2_{(k)})$;
- ou un modèle proportionnel (distribution log-normale pour $\theta_{i(k)}$) : $\theta_{i(k)} = \beta_{(k)} \times exp(b_{i(k)})$, soit $ln(\theta_{i(k)}) \sim N(ln(\beta_{(k)}), \omega^2_{(k)})$, où $\omega^2_{(k)}$ représente la variance du $k$ème effet aléatoire.

  On note d la dimension de b et $b = (b_{i(1)}, b_{i(2)}, ..., b_{i(d)})$ le vecteur des effets aléatoires chez le sujet i.
  Les effets aléatoires b_i sont supposés suivre une loi multinormale : $(b_{i(1)}, b_{i(2)}, ..., b_{i(d)}) \sim N(0, \Omega)$ où Ω est la matrice de variance-covariance $d \times d$ des paramètres aléatoires. On suppose souvent Ω diagonale.

Estimation (1/4)

Une fois le modèle écrit, comment en estimer les paramètres ?

On note Ψ l'ensemble des paramètres de population à estimer.
Ψ comprend notamment :

- les effets fixes β ;
- la variabilité inter-individuelle (les paramètres de Ω) ;
- la variabilité résiduelle (les paramètres du modèle d'erreur de g comme a, b et c) ;

Il existe deux approches pour l'estimation des paramètres :

- l'estimation par maximum de vraisemblance ;
- l'approche bayésienne.

Ces deux approches sont différentes même si en pratique les estimations obtenues se rejoignent et ces deux analyses sont asymptotiquement équivalentes.

Remarque : A moins d'utiliser un contexte bayésien, une inférence par le maximum de vraisemblance est en général utilisée.

Estimation (2/4)

Rappel sur le maximum de vraisemblance :

- Soit Y une variable aléatoire qui suit une loi normale : $Y \sim N(\mu, \sigma^2)$.
  - $y_1, ...y_n$ réalisations de Y
  - Estimation de μ et σ^2
- Vraisemblance
  - La vraisemblance au vue de l'observation y_j est définie par densité de Y en y_j (en fonction de μ et σ^2) :

$$p(y_j) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}}$$

  - Et la vraisemblance des paramètres vue les observations

$$L_y(\mu, \sigma) = \frac{1}{(\sigma\sqrt{2\Pi})^n} \Pi_{j=1}^{n} e^{-\frac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}}$$

On note l la log-vraisemblance, *i.e.* $l(\mu, \sigma) = ln(L_y(\mu, \sigma))$.

Estimation (3/4)

Suite du rappel sur le maximum de vraisemblance :

- On cherche à trouver les paramètres qui maximisent la vraisemblance.
- Estimateur du maximum de vraisemblance :
  - Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance :

    $$\begin{cases} \frac{\partial l(\hat{\theta})}{\partial \theta} = 0 \\ \frac{\partial l^2(\hat{\theta})}{\partial \theta^2} < 0 \end{cases}$$

  - L'estimateur du maximum de vraisemblance fournit directement une estimation de la variance des estimateurs.

    La variance de $\hat{\theta}$ est $Var(\hat{\theta}) = -\frac{\partial l^2(\hat{\theta})}{\partial \theta^2}$

Estimation (4/4)

- La vraisemblance dépend des paramètres θ.
- L'estimation par maximum de vraisemblance nécessite une expression des dérivées premières et secondes de la log-vraisemblance par rapport à tous les paramètres dans θ.
- Ces dérivées sont faciles à obtenir dans le cas du modèle linéaire mais pas dans le cas du modèle non linéaire.
- Si f est non linéaire en θ, la vraisemblance ne peut s'écrire de manière explicite (pas d'expression analytique).

  $\Rightarrow$ La maximisation de la vraisemblance se fait par des algorithmes basés sur des approximations.

Estimation dans le cas des modèles non linéaires

Le calcul de la vraisemblance peut s'effectuer par :

- des méthodes approchées qui permettent de linéariser le modèle et ainsi de se ramener à une expression explicite de la vraisemblance
- des méthodes exactes
  - approximation déterministe de la vraisemblance par quadrature de Gauss
  - approximation stochastique de la vraisemblance par simulations de Monte-Carlo
- ...

Estimation dans le cas des modèles non linéaires

Le calcul de la vraisemblance peut s'effectuer par :

- des méthodes approchées qui permettent de linéariser le modèle et ainsi de se ramener à une expression explicite de la vraisemblance
- des méthodes exactes
  - approximation déterministe de la vraisemblance par quadrature de Gauss
  - approximation stochastique de la vraisemblance par simulations de Monte-Carlo
- ...

Estimation dans le cas des modèles non linéaires

Le calcul de la vraisemblance peut s'effectuer par :

- des méthodes approchées qui permettent de linéariser le modèle et ainsi de se ramener à une expression explicite de la vraisemblance
- des méthodes exactes
  - approximation déterministe de la vraisemblance par quadrature de Gauss
  - approximation stochastique de la vraisemblance par simulations de Monte-Carlo
- ...

Estimation dans le cas des modèles non linéaires

Le calcul de la vraisemblance peut s'effectuer par :

- des méthodes approchées qui permettent de linéariser le modèle et ainsi de se ramener à une expression explicite de la vraisemblance
- des méthodes exactes
  - approximation déterministe de la vraisemblance par quadrature de Gauss
  - approximation stochastique de la vraisemblance par simulations de Monte-Carlo
- ...

Estimation par des logiciels

Pour l'estimation de paramètres par maximum de vraisemblance, des logiciels sont disponibles :

- NONMEM (L. Sheiner et S. Beal)
  - distributions normales ou mixtures de normales
  - plusieurs approximations : FO, FOCE, Laplace
  - algorithme SAEM
- R
  - librairie nlme (J. Pinheiro et D. Bates) implémentant la méthode FOCE
  - librairie saemix pour l'algorithme SAEM
- SAS
  - proc NLMIXED (Wolfinger) : quadrature gaussienne adaptative
  - proc MIXNLIN (Vonesh)
- P-Pharm : algorithme EM (F. Mentré et R. Gomeni)
- MONOLIX : algorithme SAEM (M. Lavielle)
- ...

Évaluation de modèles

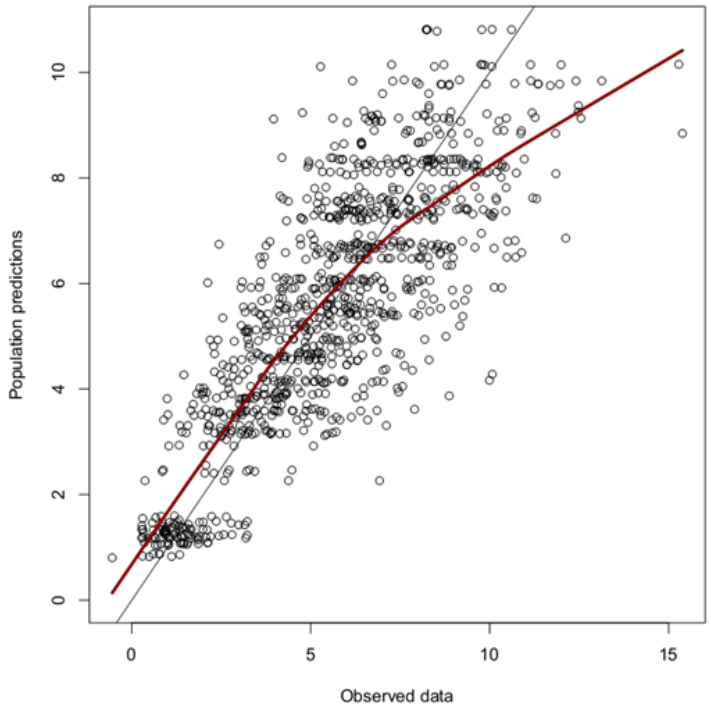
- Évaluation :
  - du modèle structurel (biologique)
  - des covariables
  - du modèle de la variabilité inter-individuelle et résiduelle

- Caractéristiques d'un modèle évalué :
  - bon ajustement aux données
  - simplicité (parcimonie)
  - capacité prédictive

Graphiques de diagnostics

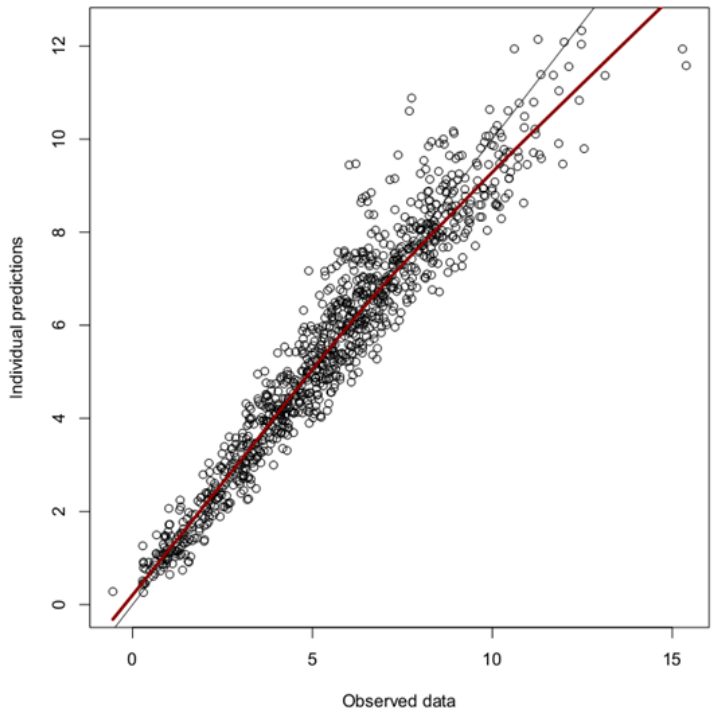
- Inclus dans les principaux logiciels (NONMEM, MONOLIX)
- Examen global (prédictions individuelles et population vs observations)
- Examen du modèle structurel et statistique
  - Examen des résidus moyens et individuels vs prédictions

Examen des prédictions en utilisant les paramètres de population



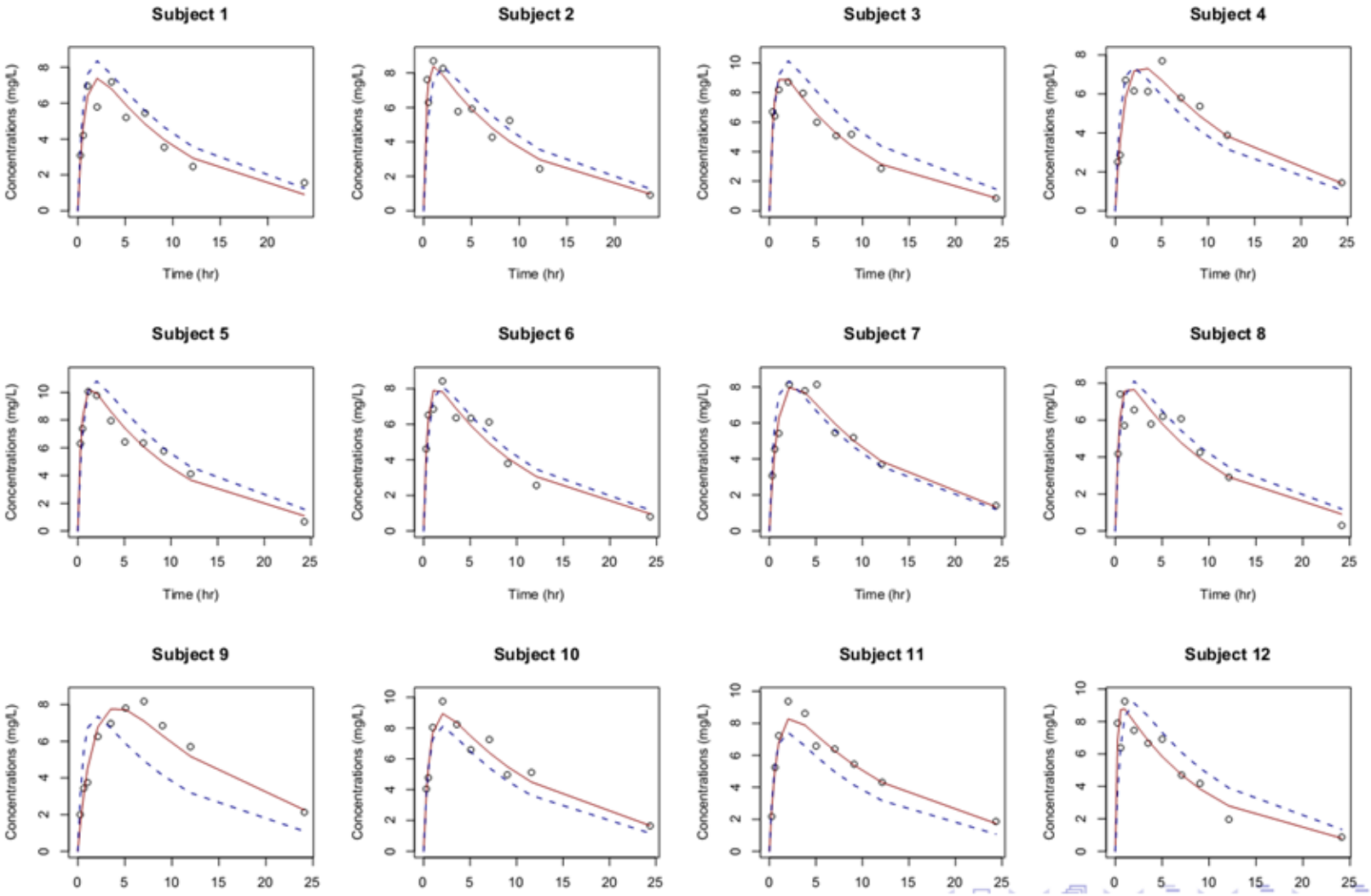
- Diagnostic général
- Pas trop de biais
- Donne une idée de la capacité du modèle à prédire de nouvelles données

Examen des résidus des paramètres individuels

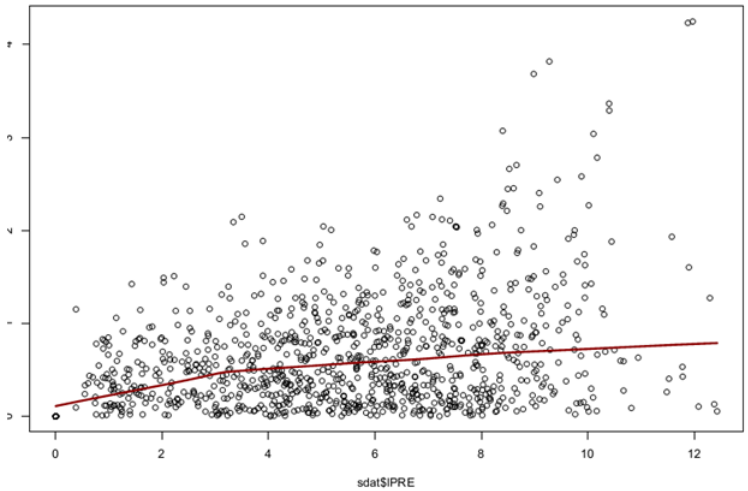


- Graphiques plus adaptés en cas de données riches par patient
- Évalue le modèle structurel

Ajustements individuels



Exemple des résidus individuels en fonction des prédictions individuelles pour détecter un modèle d'erreur mal adapté



- Modèle additif utilisé au lieu de proportionnel
- Tendance nette dans les résidus
- La variance de l'erreur augmente avec la prédiction

Démarche d'analyse (1/6)

Démarche pour construire son analyse :

- Connaissance du système (et du médicament)
- Examen des données
- Construction d'un modèle approprié
- Estimation des paramètres et de leur écart-type
- Examen des graphiques d'ajustement
- Tests statistiques en particulier pour la comparaison de modèle

Démarche d'analyse (2/6)

Examen préalable des données

- Toujours examiner les graphiques exploratoires
- Surveiller d'éventuels points aberrants
- Si possible, tester la capacité de modèles candidats à ajuster les données individuelles (sous R par exemple)

Démarche d'analyse (3/6)

Choix d'un modèle initial

- Modèle structurel (type d'absorption, modèle de l'effet du traitement,...)
- Choix des paramètres initiaux dans l'algorithme d'estimation : issu de connaissances *a priori* (littérature) ou de l'ajustement individuels de quelques sujets
- Modélisation de la variabilité individuelle (modèle additif $\theta_i = \beta + b_i$ ou exponentiel $\theta_i = \beta \times exp(b_i)$) : lorsque les paramètres sont positifs, privilégier un modèle exponentiel

Démarche d'analyse (4/6)

Estimation des paramètres et évaluation des écarts-types (SE pour standard errors en anglais)

- Critère de qualité du modèle
- Obtenir des SE (ou SE relatives RSE) petites est une indication que les paramètres sont bien estimés
- Qu'est ce qu'une bonne précision d'estimation ? A la louche :
  - $< 20 - 30$ % pour un effet fixe
  - $< 30 - 50$ % pour les paramètres de variabilité inter-individuelle ω^2
  - $< 10 - 30$ % pour la variance résiduelle

Démarche d'analyse (5/6)

Évaluation de modèles et examen des graphiques d'ajustement (déjà détaillés précédemment)

Démarche d'analyse (6/6)

Tests d'hypothèses : tests du rapport de vraisemblance

- Test du rapport de vraisemblance :
  - Compare deux modèles emboités, M1 à p paramètres et M2 à $p+q$ paramètres
  - Test H_0 : les q paramètres supplémentaires de M2 sont égaux à 0
    Statistique de test $T = -2(ln(L2) - ln(L1)) \sim \chi^2(q)$ (asymptotiquement)
- Modèles non emboités
  - AIC, BIC : attention ce ne sont pas des tests, ne fournit pas de p-value
  - AIC= $-2ln(L) + 2p$
  - BIC= $-2ln(L) + pln(n_{tot})$
  - BIC souvent préférable

Conclusion

- Les modèles à effets mixtes permettent l'estimation précise des paramètres moyens d'un échantillon
- même quand l'estimation au niveau individuel n'est pas faisable (pas assez de données)
- Méthodes d'estimation
  - Principale difficulté lorsque le modèle biologique est non linéaire