Master Agronomie, AgroParisTech, 2014

Modèle non linéaire

David Makowski
INRA

Rappels sur le modèle linéaire

$$y = \underbrace{\theta_0 + \theta_1 x_1 + \ldots + \theta_p x_p}_{} + \varepsilon$$

Variable continue

Combinaison linéaire de paramètres et de variables explicatives continues ou discrètes

Résidu

Rappels sur le modèle linéaire

$$y = \theta_0 + \theta_1\, x_1 + \ldots + \theta_p\, x_p + \varepsilon$$

On suppose $E(\varepsilon) = 0$

Conséquence: $E(y) = \theta_0 + \theta_1\, x_1 + \ldots + \theta_p\, x_p$

La réponse moyenne est une *combinaison linéaire* des variables explicatives et des paramètres.

Rappels sur le modèle linéaire

Exemples de modèles linéaires:

- Régression linéaire simple.
- Régression linéaire multiple.
- ANOVA.
- Analyse de covariance.

Deux limites du modèle linéaire

- **Ne permet pas de modéliser des variables de réponse discrètes (ex: $y=0$ ou $y=1$, comptage…) .**

- **Ne permet pas de tenir compte d'une relation non linéaire entre la variable de réponse y et les variables explicatives $x_1, \ldots, x_p$.**

Deux extensions

- Le modèle linéaire généralisé
- Le modèle non linéaire

Modèle non linéaire

- Permet d'utiliser une relation non linéaire quelconque entre la variable de réponse y et des variables explicatives x_1-x_p.

- Description plus réaliste des phénomènes.

- Utilisation de paramètres ayant une signification intéressante.

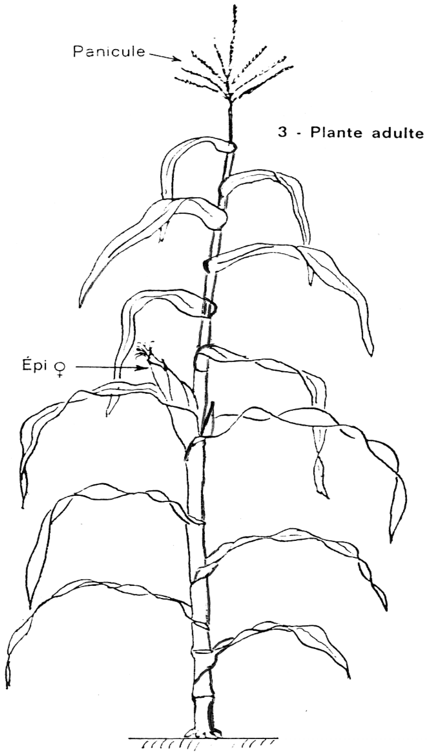
- Calculs plus complexes.

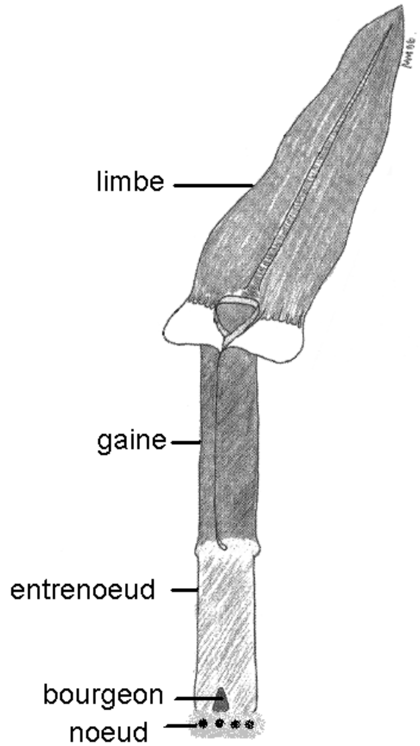
Quatre problèmes liés à l'estimation des paramètres d'un modèle non linéaire

1. Quels paramètres estimer ?	Paramétrage.
2. Quelle information utiliser ?	Plan d'expérience, répétitions.
3. Quelle méthode d'estimation utiliser ?	Moindres carrés, max. vraisemblance.
4. Précision des estimateurs ?	Variances, corrélations, test.

Exemple « Maïs »

« Modèle qui simule la longueur du limbe en fonction de la somme de température ».





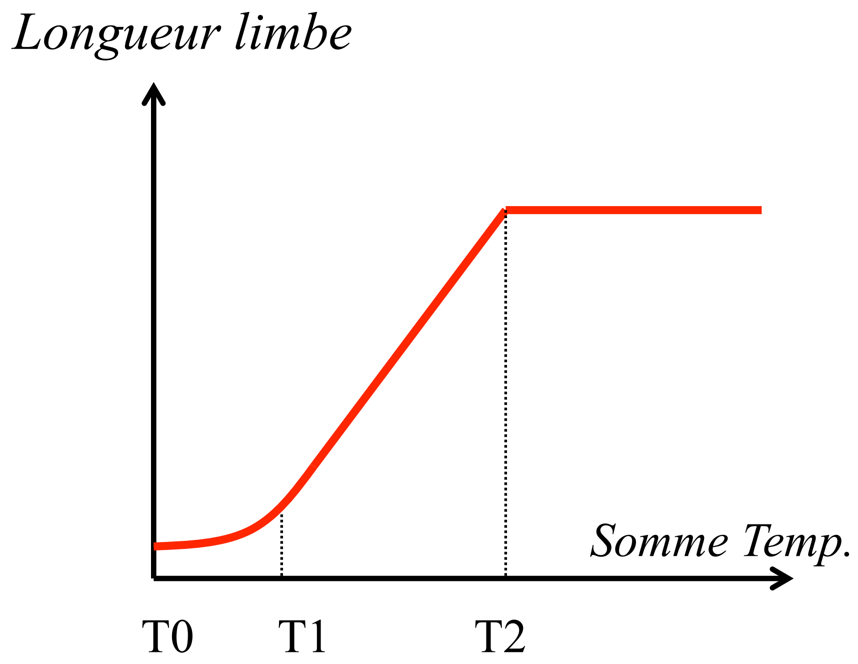
Représentation d'un plan de maïs et d'un phytomère
(adapté d'après Scanlon et Freeling 1997).

Le modèle (pour un phytomère)

$$f(x;\theta) = L_{MIN} e^{R_1(x-T_0)} \qquad T_0 < x \leq T_1$$

$$f(x;\theta) = \alpha + \beta(x - T_1) \qquad T_1 < x \leq T_2$$

$$f(x;\theta) = L_{MAX} \qquad T_2 < x$$



Quels paramètres doit-on estimer ?

Quels paramètres doit-on estimer ?

- *Contraintes pour assurer la continuité de la fonction:*

$$\alpha = L_{MIN} e^{R_1 (T_1 - T_0)} \qquad L_{MAX} = \alpha + \beta (T_2 - T_1)$$

- *Contrainte pour assurer la continuité de la dérivé entre la première et la deuxième phase:*

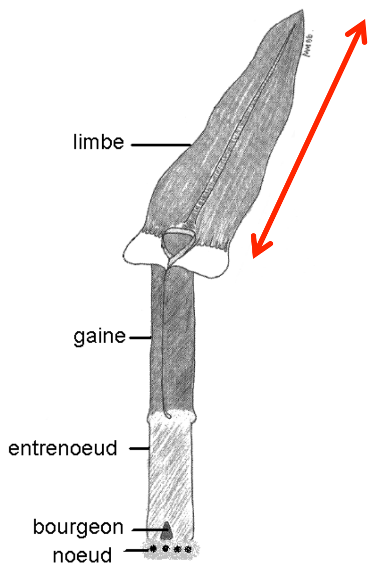
$$\beta = L_{MIN} R_1 e^{R_1 (T_1 - T_0)}$$

- *On peut fixer L_{MIN} à une valeur initiale.*

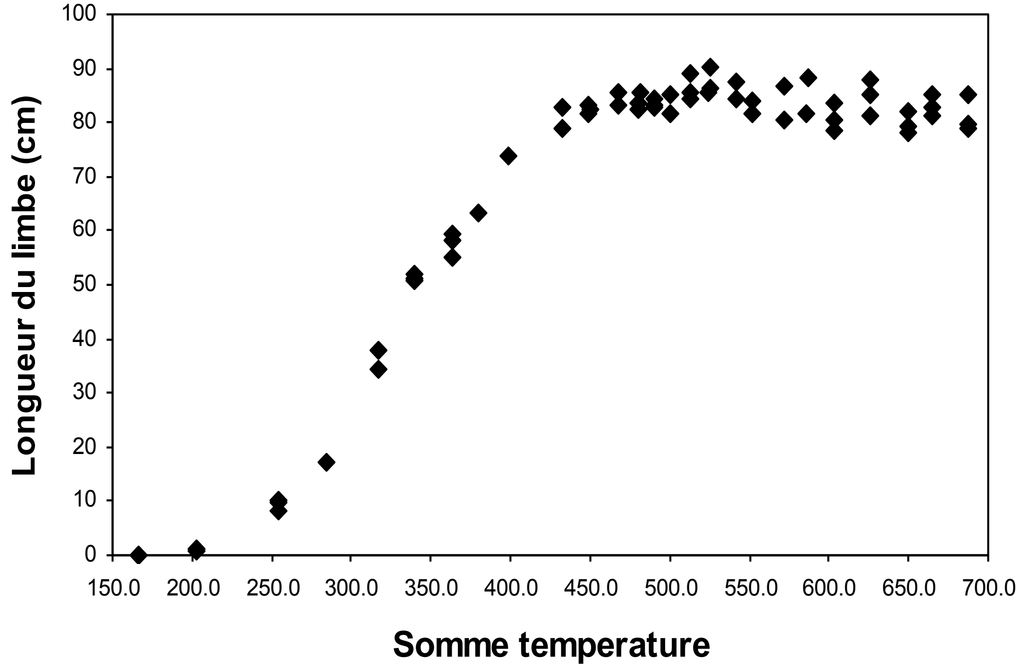
→ **Quatre paramètres à estimer : R_1, T_0, T_1, T_2 .**

Quelle information utiliser ?

- Mesures de longueurs de limbe obtenues sur une parcelle pour le 9ème phytomère.
- Mesures réalisées 2 à 3 fois par semaine.
- Une à trois mesures obtenues pour chaque date sur différentes plantes.



Phytomère 9



Quelle méthode d'estimation utiliser ?

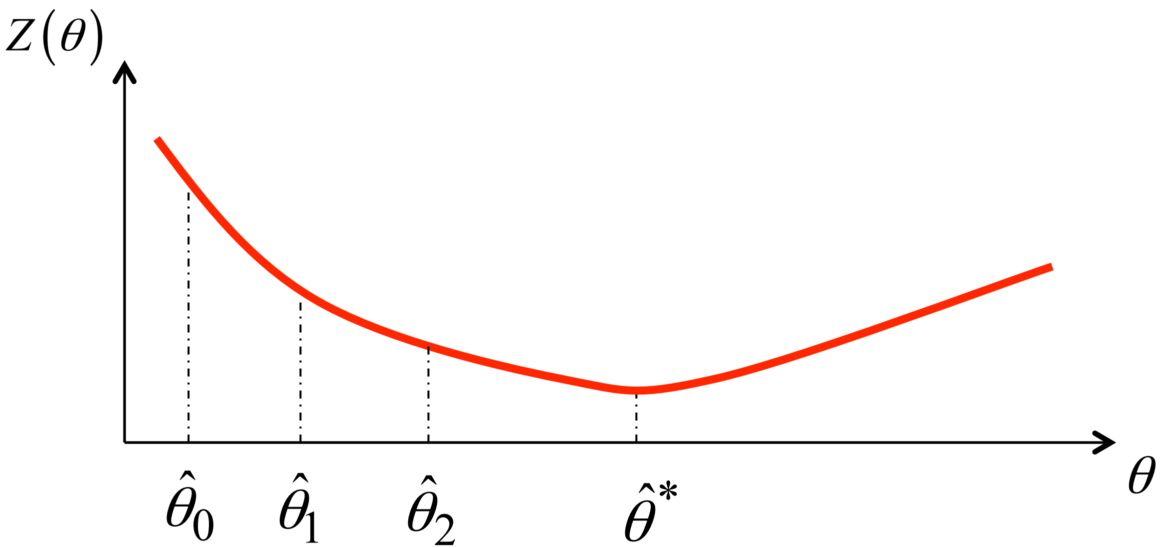
1er possibilité : La méthode des moindres carrés ordinaires

Trouver la valeur de θ qui minimise : $$Z(\theta) = \sum_{i=1}^{N} \left[y_i - f(x_i;\theta) \right]^2$$

Problème :

- le modèle est non linéaire,
- on ne peut pas trouver l'expression analytique des estimateurs.

Appliquer la méthode des MCO avec un algorithme itératif



L'algorithme de Gauss-Newton

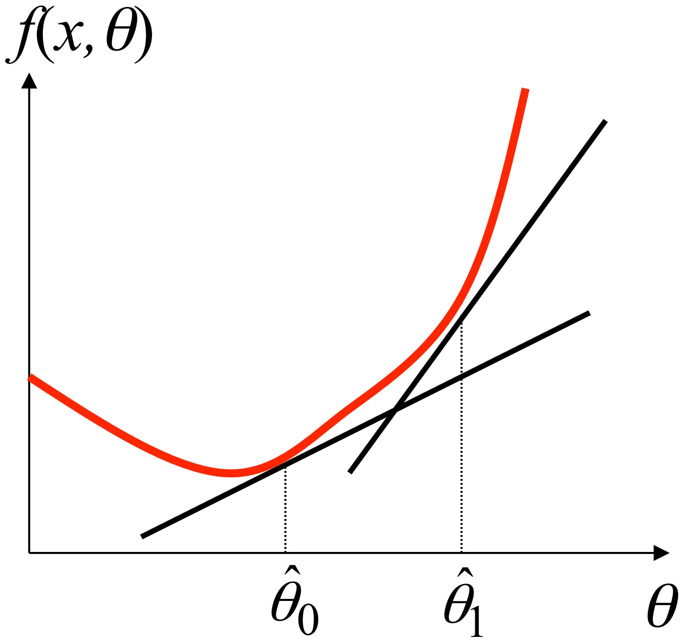
1. On définit une valeur initiale $\hat{\theta}_0$

2. On linéarise le modèle par un dév. de Taylor

$$f(x;\theta) \approx f\left(x;\hat{\theta}_0\right) + \sum_{j=1}^{p} \left.\frac{\partial f(x;\theta)}{\partial \theta_j}\right|_{\hat{\theta}_0} \left(\theta_j - \hat{\theta}_{0j}\right)$$

3. On calcule l'estimateur des moindres carrés avec le modèle linéarisé → $\hat{\theta}_1$

4. Retour à l'étape 1 en remplaçant $\hat{\theta}_0$ par $\hat{\theta}_1$.



Arrêt si $\sum_{i=1}^{N}\left[y_i - f\left(x_i;\hat{\theta}_{k+1}\right)\right]^2 - \sum_{i=1}^{N}\left[y_i - f\left(x_i;\hat{\theta}_k\right)\right]^2$ est négligeable.

L'algorithme de Gauss-Newton

Questions

$$f\left(x;\theta\right) = e^{\theta x}$$

- Linéariser le modèle à l'aide d'un développement de Taylor

$$f\left(x;\theta\right) \approx f\left(x;\hat{\theta}_0\right) + \left.\frac{df\left(x;\theta\right)}{d\theta}\right|_{\hat{\theta}_0} \left(\theta - \hat{\theta}_0\right)$$

- Exprimer le modèle sous la forme : $A + B\ \theta$

L'algorithme de Gauss-Newton

$$f\left(x;\theta\right) = e^{\theta x}$$

$$e^{\theta x} \approx e^{\hat{\theta}_0 x} + \left(\theta - \hat{\theta}_0\right) x e^{\hat{\theta}_0 x}$$

$$e^{\theta x} \approx e^{\hat{\theta}_0 x} - \hat{\theta}_0 x e^{\hat{\theta}_0 x} + x e^{\hat{\theta}_0 x} \times \theta$$

Estimateur: $$\hat{\theta}_1 = \hat{\theta}_0 + \frac{\sum_{i=1}^{N} x_i e^{\hat{\theta}_0 x_i}\left(Y_i - e^{\hat{\theta}_0 x_i}\right)}{\sum_{i=1}^{N} x_i^2 e^{2\hat{\theta}_0 x_i}}$$

L’algorithme de Gauss-Newton

Aspects pratiques

- On utilise un logiciel statistique (SAS, S+, R, MatLab…)

- On donne en entrée:

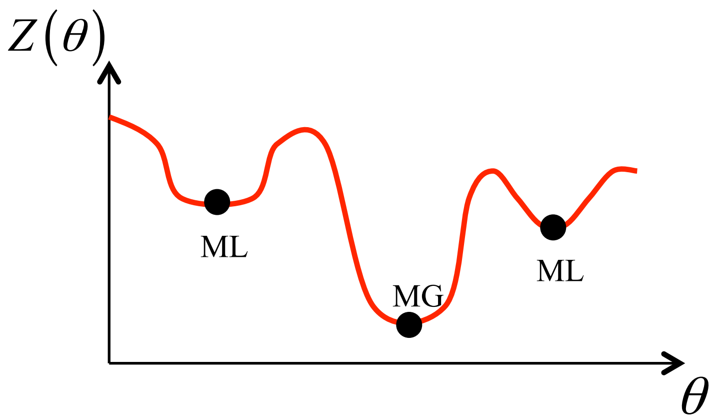
    - des données,

    - un modèle,

    - des valeurs initiales des paramètres.

- Le logiciel fournit en sortie les valeurs estimées des paramètres.

Minimum locaux et minimum globaux



→ **Essayez plusieurs valeurs initiales !**

Quelle méthode d'estimation utiliser ?

1[er] possibilité : La méthode des moindres carrés ordinaires

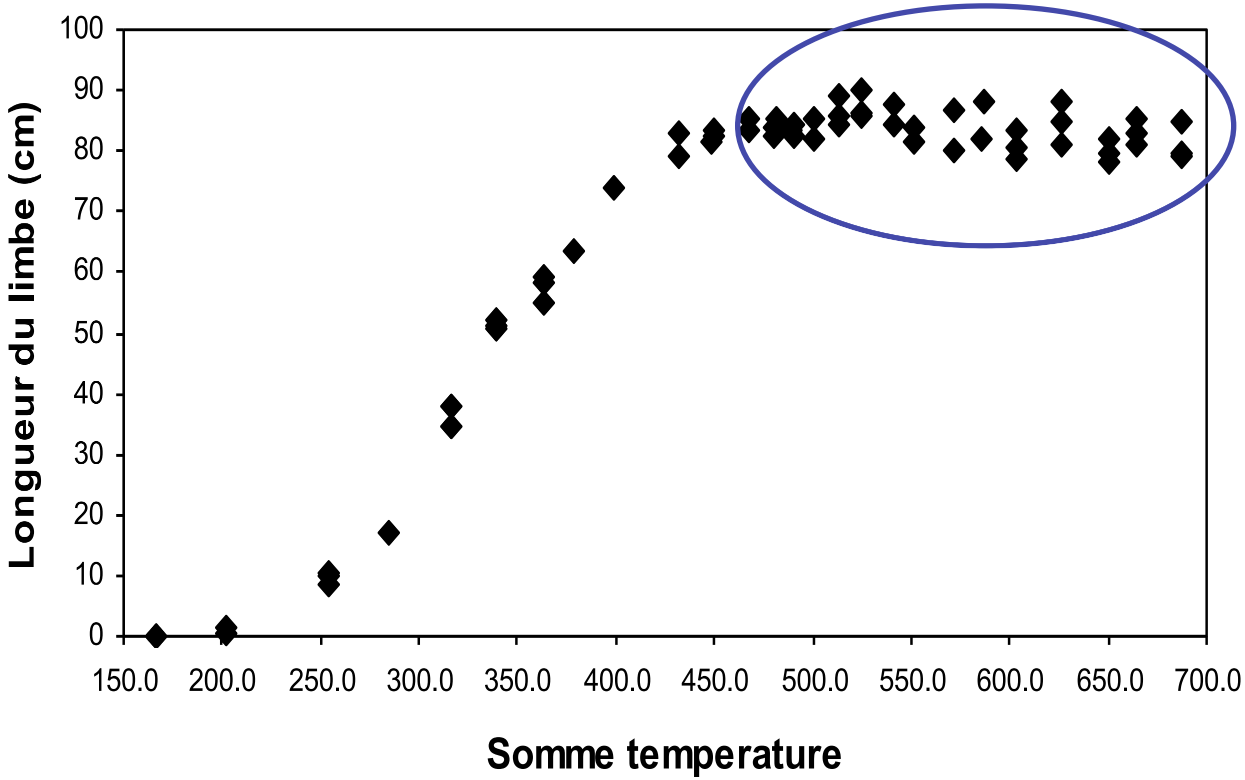
Trouver la valeur de θ qui minimise : $Z(\theta) = \sum_{i=1}^{N} \left[y_i - f(x_i;\theta) \right]^2$

Inconvénient :

- Les estimateurs ne sont pas de variances minimales si les résidus ont des variances hétérogènes.

- Or ici, les erreurs de mesures sont plus grandes pour les limbes de grandes tailles.

Variance plus grande ici.



Quelle méthode d'estimation utiliser ?

La méthode des moindres carrés pondérés

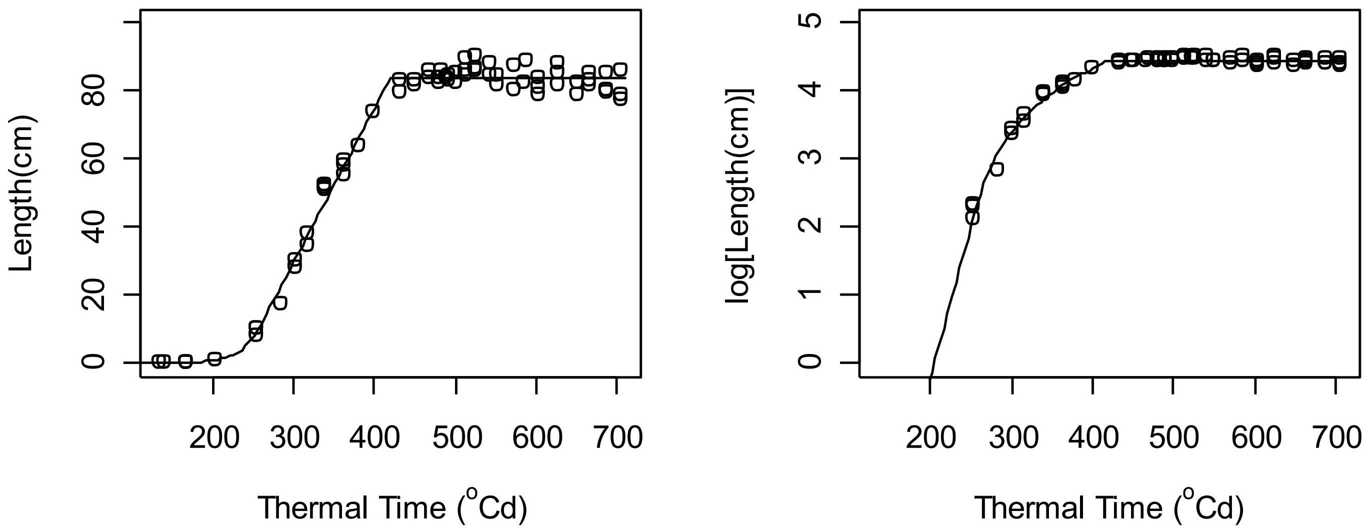
Trouver la valeur de θ qui minimise : $Z(\theta) = \sum_{i=1}^{N} \frac{\left[y_i - f(x_i;\theta)\right]^2}{\sigma_i^2}$

avec $\sigma_i^2 = \frac{1}{K-1} \sum_{k=1}^{K} \left(y_{ik} - \bar{y}_{i\cdot}\right)^2$

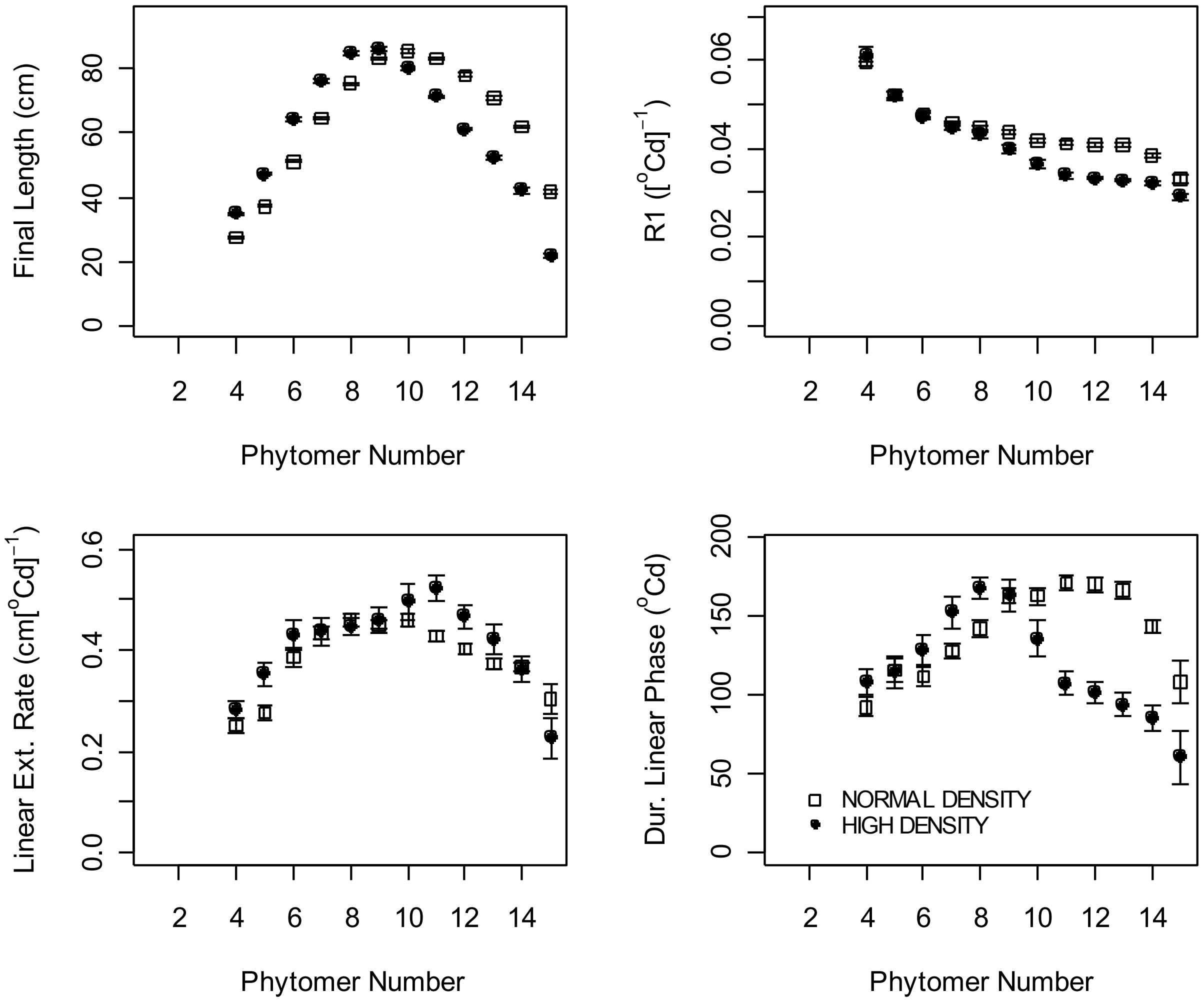
Définire la variance des résidus comme une fonction croissante de la longueur du limbe

$$\mathrm{var}\left[y_i - f(x_i;\theta)\right] = \sigma^2 f(x_i;\theta)^\tau$$

On estime θ, σ et τ à partir des données.


Length(cm)
0
20
40
60
80
200
300
400
500
600
700
Thermal Time (°Cd)
log[Length(cm)]
0
1
2
3
4
5
200
300
400
500
600
700
Thermal Time (°Cd)

Ces estimateurs sont-ils précis ?



Utilisation de R pour estimer les paramètres d’un modèle non linéaire

```
Fit<-nls(y~FONCTION(x, Theta0, Theta1, Theta2),
    start=list(Theta0=9, Theta1=0.04, Theta2=100),
    data=TAB)
summary(Fit)
coef(Fit)
```

Exemple « fertilisation du blé »

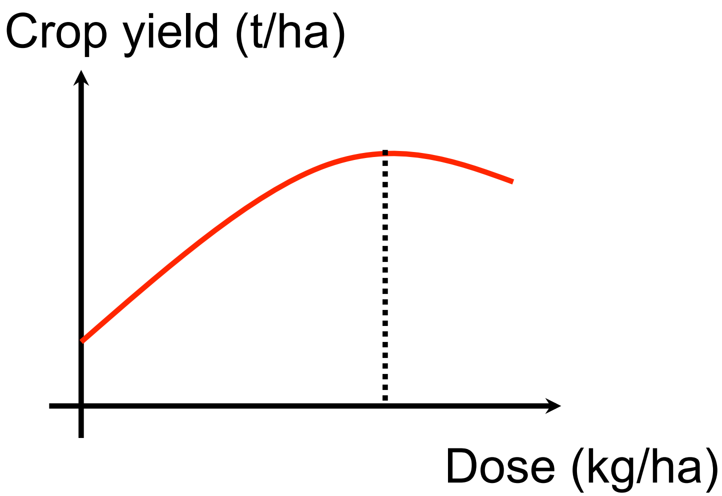
Quelle dose d’engrais N appliquer ?

2. Modèle non linéaire

Exemples de modèles pour optimiser la dose d'engrais

Model linéaire

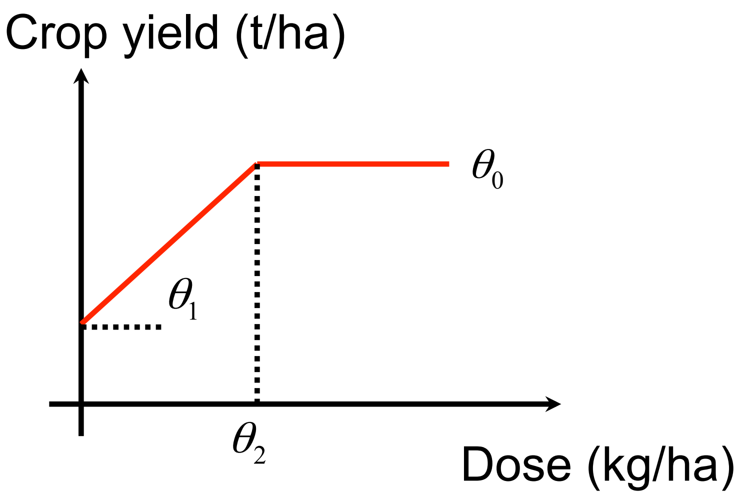
$$z = \theta_0 + \theta_1 x + \theta_2 x^2$$



Model non linéaire

$$z = \theta_0 \ \text{if}\, x \geq \theta_2$$

$$z = \theta_0 + \theta_1 \left(x - \theta_2 \right) \ \text{if}\, x < \theta_2$$



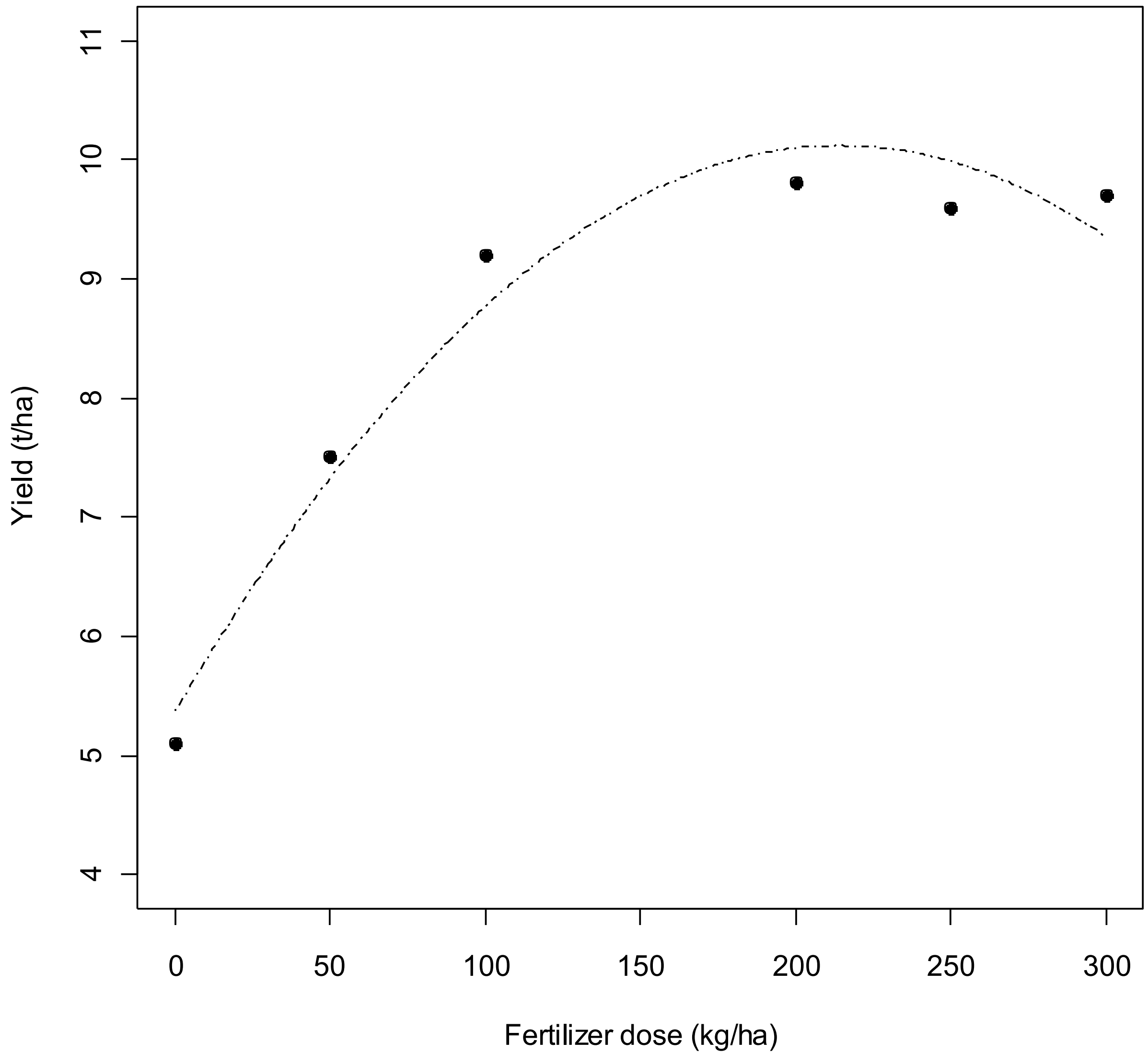
Les données

```
x <- c(0, 50, 100, 200, 250, 300)        #Values of x (fertilizer dose)
y<- c(5.1, 7.5, 9.2, 9.8, 9.6, 9.7)      #Values of y (yield)
TAB<-data.frame(x, y)                    #dataset TAB
TAB                                      #print the dataset TAB
TAB$x                                    #column x of TAB
TAB$y                                    #column y of TAB
x                                        #column x of TAB
y                                        #column y of TAB
```

Estimation des paramètres – modèle linéaire

```
x2<-x*x                         # New variable
Fit<-lm(y~x+x2)                 # Parameter estimation by least squares
                                # for the quadratic model

summary(Fit)                    # Results
coef(Fit)                       # The three estimated parameter values
Parameters<-coef(Fit)

X11()                           #New window
par(mfrow=c(1,1))

plot(x,y, xlab="Fertilizer dose (kg/ha)", ylab="Yield (t/ha)", pch=19, ylim=c(4,11))

Pred<-Parameters[1]+Parameters[2]*(0:300)+Parameters[3]*(0:300)^2

lines(0:300, Pred, lty=4)
```

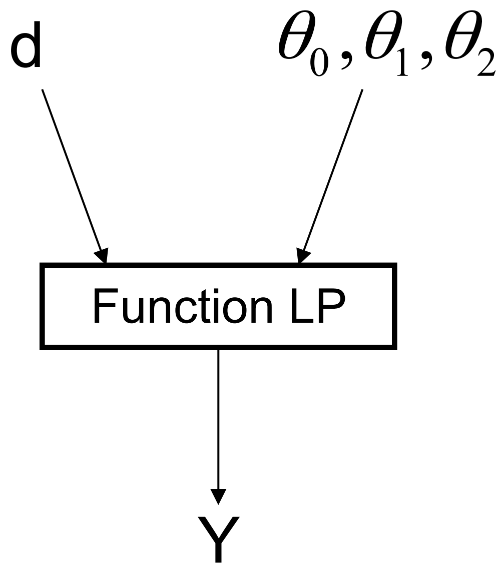
2. Modèle non linéaire



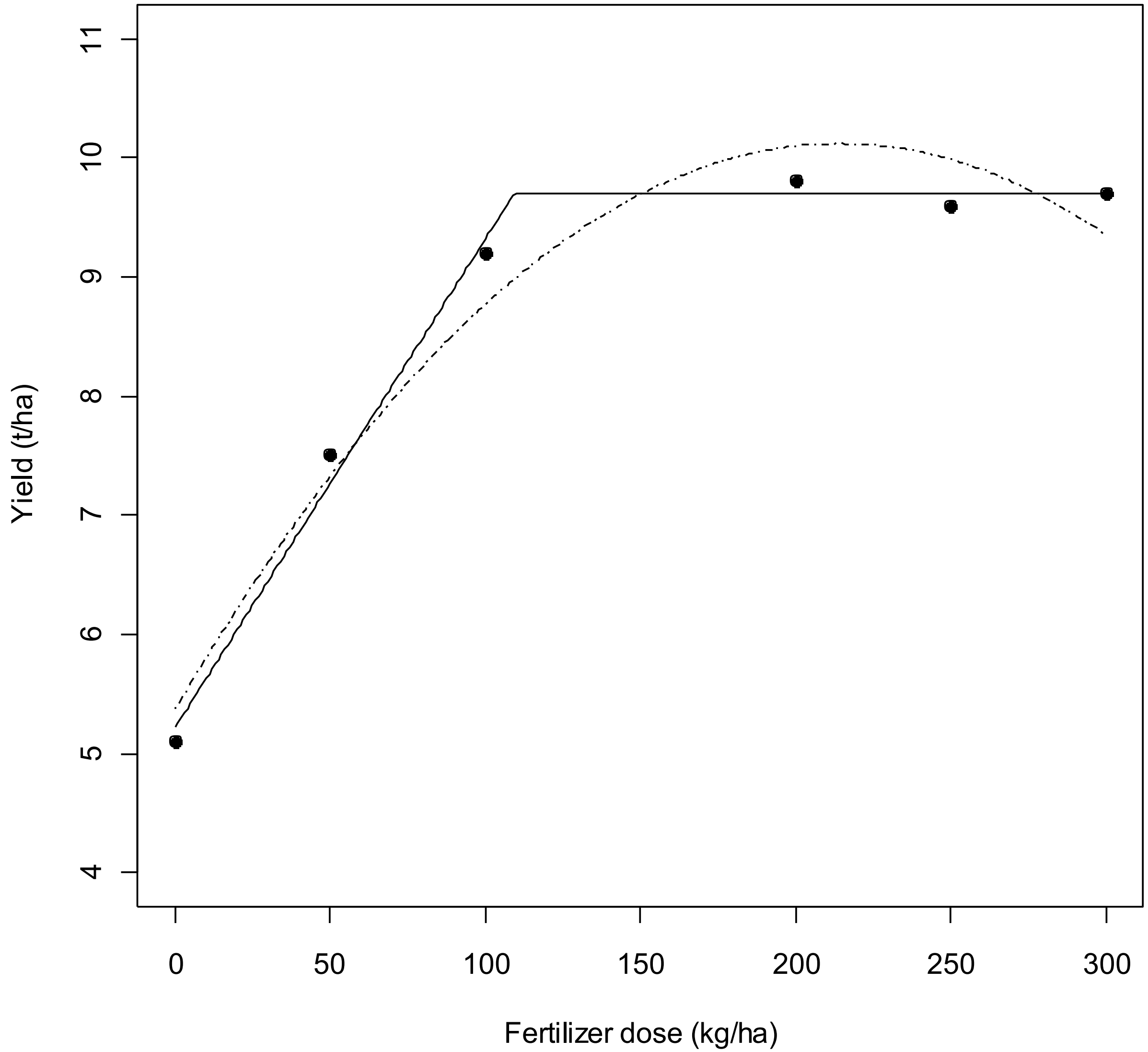
Estimation des paramètres – modèle non linéaire

```
LP<-function(d, Theta0, Theta1, Theta2) {
        Y<-Theta0+Theta1*(d-Theta2)
        Y[d>=Theta2]<-Theta0
        return(Y)
        }

Fit<-nls(y~LP(x, Theta0, Theta1, Theta2), start=list(Theta0=9, Theta1=0.04,
Theta2=100), data=TAB)

summary(Fit)

Parameters<-coef(Fit)

Pred<-Parameters[1]+Parameters[2]*(0:300-Parameters[3])
Pred[Pred>Parameters[1]]<-Parameters[1]

lines(0:300, Pred)
```

2. Modèle non linéaire



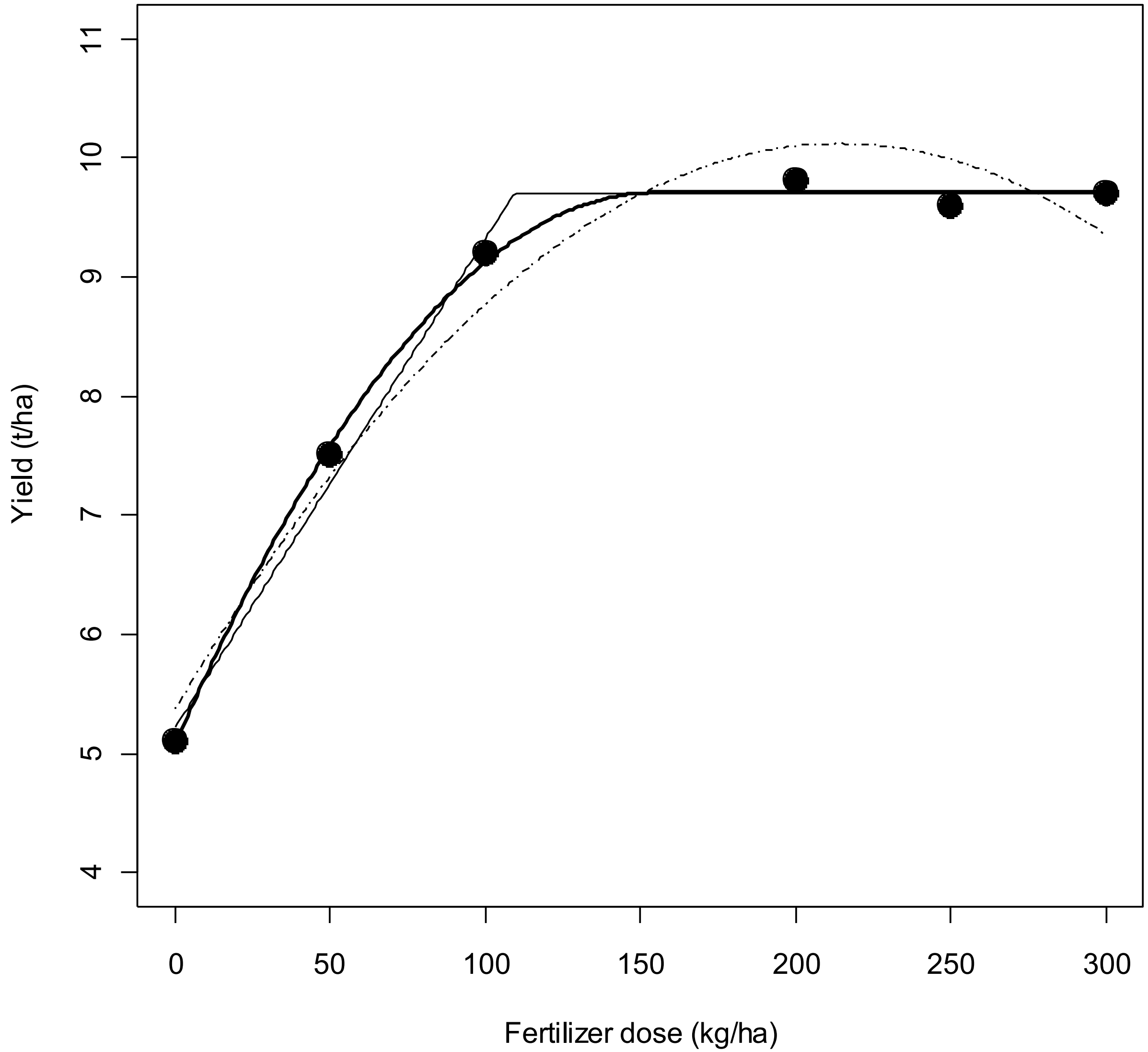
2. Modèle non linéaire



2. Modèle non linéaire

```
QP<-function(d, Theta0, Theta1, Theta2) {
        Y<-Theta0+Theta1*(d-Theta2)^2
        Y[d>=Theta2]<-Theta0
        return(Y)
        }

Fit<-nls(y~QP(x, Theta0, Theta1, Theta2), start=list(Theta0=9,
Theta1=-0.004, Theta2=100), data=TAB)

summary(Fit)

Parameters<-coef(Fit)

Pred<-Parameters[1]+Parameters[2]*(0:300-Parameters[3])^2
Pred[0:300>Parameters[3]]<-Parameters[1]

lines(0:300, Pred, lwd=2)
```

2. Modèle non linéaire



Conclusions

• **Le modèle linéaire ne permet pas de traiter tous les problèmes pratiques.**

• **Le modèle linéaire généralisé permet de traiter le cas des variables de réponse discrètes, notamment binaires.**

• **Le modèle non linéaire permet souvent de décrire de façon plus réaliste un phénomène physique ou biologique.**

• **Les paramètres de ces modèles peuvent être estimés à l’aide de logiciel statistiques.**

• **L’estimation des paramètres des modèles non linéaires est parfois délicate à mettre en œuvre.**

Quelques références

Agresti A. 1990. *Categorical data analysis.* Wiley

Makowski D, Monod H. 2011. *Analyse statistique des risques agro-environnementaux*. Springer

McCulloch C.E., Searle S.R. 2001. *Generalized, linear, and mixed models*. Wiley

Seber GAF, Wild CJ. 2003. *Nonlinear regression*. Wiley.

Wallach D, Makowski D, Jones JW. 2006. *Working with dynamic crop models*. Elsevier.

0653 couv agro ok 29/06/2011 13:40 Page 1

Collection
Statistique
et probabilités
appliquées

Cette collection met à la disposition du public intéressé par la statistique (étudiants, enseignants, chercheurs) des ouvrages qui concilient effort pédagogique et travail permanent de mise à jour. Cette démarche implique de prendre en compte de façon sélective et critique les renouvellements des concepts, des champs d'application et des outils de traitement. Seules une compréhension profonde et une appropriation des connaissances permettront de s'adapter aux évolutions qui n'ont pas fini de bouleverser cette discipline.

David Makowski, Hervé Monod

Analyse statistique des risques agro-environnementaux

Études de cas

Conçu comme un véritable manuel pratique, ce livre est une introduction aux méthodes statistiques les plus couramment utilisées pour l'analyse des risques agro-environnementaux. Celles-ci peuvent être regroupées au sein de trois grandes sections :

- La modélisation des risques en fonction de facteurs environnementaux et anthropiques (modèle linéaire, modèle linéaire généralisé, modèle non linéaire, modèle hiérarchique, régression quantile) ;
- L'optimisation de décisions ou de règles de décision pour mieux gérer les risques, en intégrant des variables décisionnelles dans les modèles (optimisation de seuils de décision, optimisation par simulation, analyses ROC) ;
- L'analyse et la communication des incertitudes associées aux modèles (estimation et description de distributions de probabilité, assimilation de données, analyse de sensibilité).

L'utilisation de chaque méthode est illustrée par une ou plusieurs applications à des problèmes concrets (pollution de l'eau par les nitrates, invasion par des espèces nuisibles, flux de gènes d'une culture OGM vers une culture non OGM, etc.). Les programmes informatiques R ou WinBUGS utilisés dans les exemples sont présentés et commentés en détail. A la fin de chaque chapitre, des exercices permettront aux lecteurs de tester leur compréhension des méthodes étudiées.

00 € TTC

ISBN : 978-2-8178-0250-3

9 782817 802503

springer.com

Makowski • Monod

Analyse statistique des risques agro-environnementaux

Collection
Statistique
et probabilités
appliquées

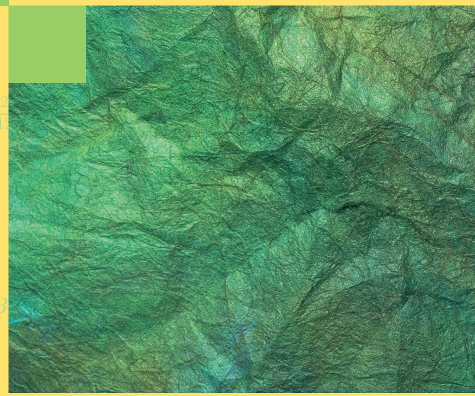
David Makowski, Hervé Monod

Analyse statistique des risques agro-environnementaux

Études de cas

$\phi(n) = \left(1 - \frac{1}{12}\right)\left(1 - \frac{2}{12}\right.$

$P(A) = \frac{25}{216}$ et $P(B$



$(\forall B \in \mathcal{B}_{\mathbb{R}})\ P_X(B) = P(X^{-1}(B))$
$= P(\{\omega \in \Omega \mid X(\omega) \in B\})$

Springer

$P(A_2) = 1 - P(\overline{A_2}) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0,491.$