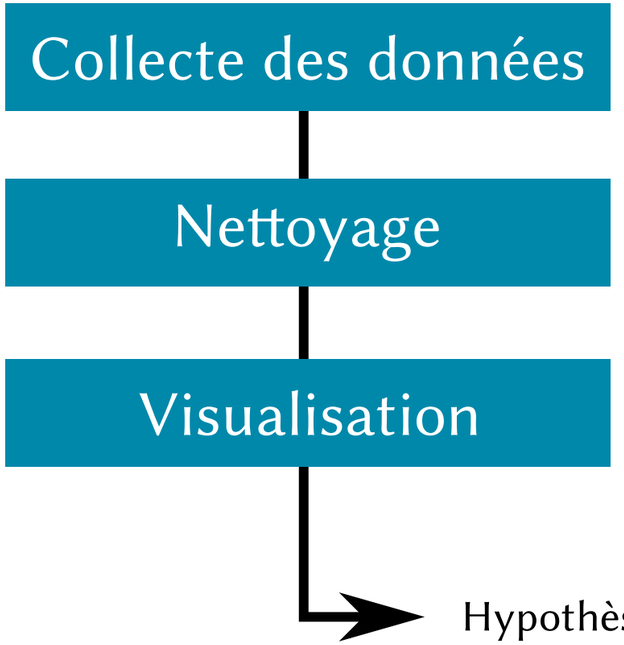
Ateliers R^3

Session 6 - Modèles linéaires ~~et modèles mixtes~~

Au menu aujourd'hui:



Hypothèses quantitatives:

« Si H est vrai alors, y devrait augmenter quand x augmente »

« Si H est vrai alors on devrait avoir une relation en cloche entre x et y »

« L'effet de x1 doit être plus grand que l'effet de x2 sur y »

« Passé une valeur seuil de x1, toutes les valeurs de y doivent être nulles »

« L'effet de x1 sur y doit être modulé par x2 »

...

Au menu aujourd'hui:

Collecte des données

Nettoyage

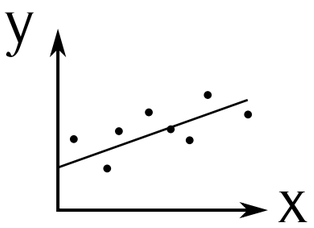
Visualisation

Hypothèses quantitatives:

Modèle statistique:

$y = a * x + b + erreur$

Si H vrai , alors $a > 0$



« Si H est vrai alors, y devrait augmenter quand x augmente »

« Si H est vrai alors on devrait avoir une relation en cloche entre x et y »

« L'effet de x1 doit être plus grand que l'effet de x2 sur y »

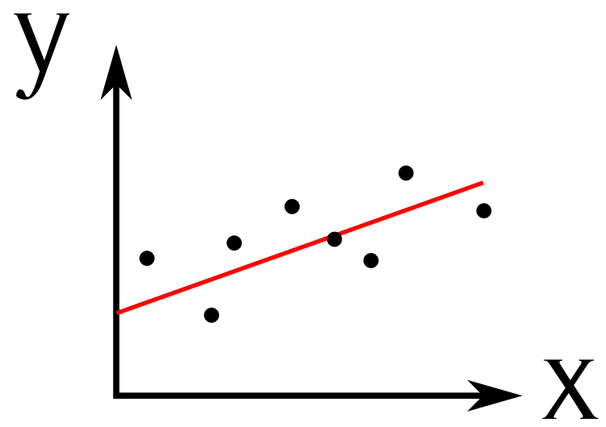
« Passé une valeur seuil de x1, toutes les valeurs de y doivent être nulles »

« L'effet de x1 sur y doit être modulé par x2 »

...

« Si H est vrai alors, y devrait augmenter quand x augmente »

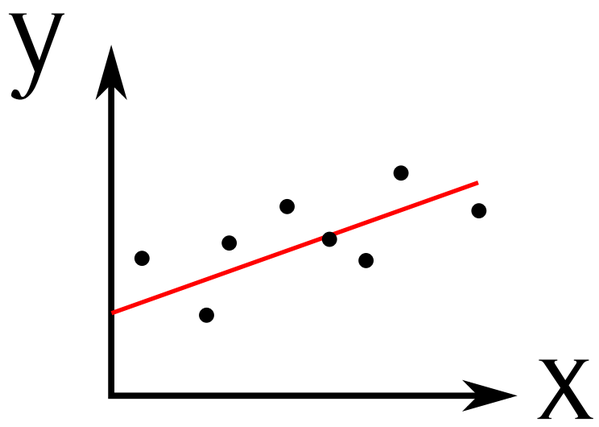
« Si mon chien aime les croquettes, alors les frétillements de la queue devraient augmenter avec la quantité de croquette »



$$y = a * x + b + \text{erreur}$$

« Si H est vrai alors, y devrait augmenter quand x augmente »

« Si mon chien aime les croquettes, alors les frétillements de la queue devraient augmenter avec la quantité de croquette »



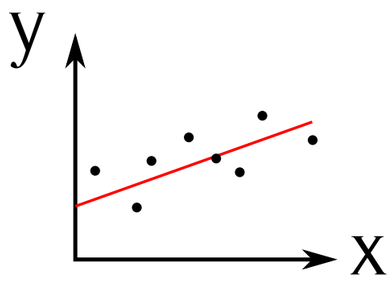
$$y = \underbrace{a * x + b}_{\text{En moyenne on observe ça pour une valeur de x}} + \underbrace{\text{erreur}}_{\text{En pratique on observe ça}}$$

« Si H est vraie, alors la pente de la droite devrait être positive (a > 0) »

« Si H est vrai alors, y devrait augmenter quand x augmente »

« Si mon chien aime les croquettes, alors les frétillements de la queue devraient augmenter avec la quantité de croquette »

Option 1

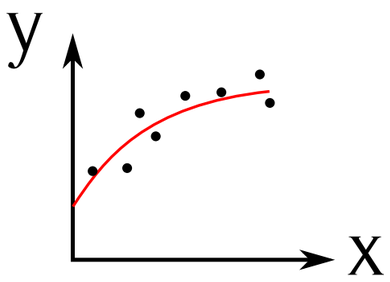


Modèle statistique:

$y = a * x + b + erreur$

Si H vrai , alors $a > 0$

Option 2

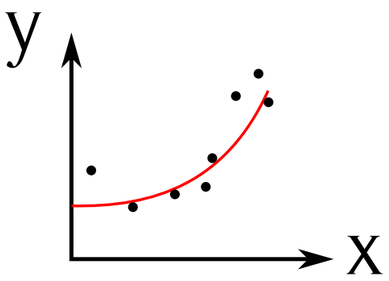


Modèle statistique:

$y = a * \sqrt{x} + b + erreur$

Si H vrai , alors $a > 0$

Option 3



Modèle statistique:

$y = a * x + b * x^2 + c + erreur$

Si H vrai , alors $b > 0$ (?)

Il faut visualiser les données !

Modèles linéaires :

$$y = a\,x + b + \text{erreur}$$
$$y = a\,\sqrt{x} + b + \text{erreur}$$
$$y = a\,x + b\,x^2 + c + \text{erreur}$$
$$y = a\,x1 + b\,x2 + c + \text{erreur}$$

Formule générale

$$y = a + b\,x1 + c\,x2 + d\,x3 + \ldots + \text{erreur}$$

les x1, x2 peuvent être des transformations de variables !

Modèles linéaires :

```
y = a x    + b          + erreur
y = a √x   + b          + erreur
y = a x    + b x² + c   + erreur
y = a x1   + b x2 + c   + erreur
```

"Variable explicative"
"Prédicteur"
"Explanatory variable"
"Predictor"

Formule générale

$$y = a + b\,x1 + c\,x2 + d\,x3 + \ldots + erreur$$

"Variable réponse"
"Response variable"

les x1, x2 peuvent être des transformations de variables !

Modèles linéaires :

$$y = a\,x + b + \text{erreur}$$
$$y = a\,\sqrt{x} + b + \text{erreur}$$
$$y = a\,x + b\,x^2 + c + \text{erreur}$$
$$y = a\,x_1 + b\,x_2 + c + \text{erreur}$$

Modèles pas linéaires :

$$y = 1 / (a + b * x_1)$$
$$y = a - \exp(-b * x_1)$$
$$y = \sin(a * x_1)$$

etc...

Estimation des paramètres

$$y = a + b\ x1 + c\ x2 + d\ x3 + \ldots + \text{erreur}$$

On a des observations de chaque variable

y	x_1	x_2	x_3
30.5	-1.1	-0.1	0.2
16.7	-2.5	-2.4	-2.8
18.3	-3.1	1.1	-1.8
29.1	1.4	1.8	0.8

Quelles sont les valeurs de a, b, c et d les plus *vraisemblables* étant donné ces observations ?

"celles qui minimisent l'écart entre la prédiction et l'observé"

Estimation des paramètres

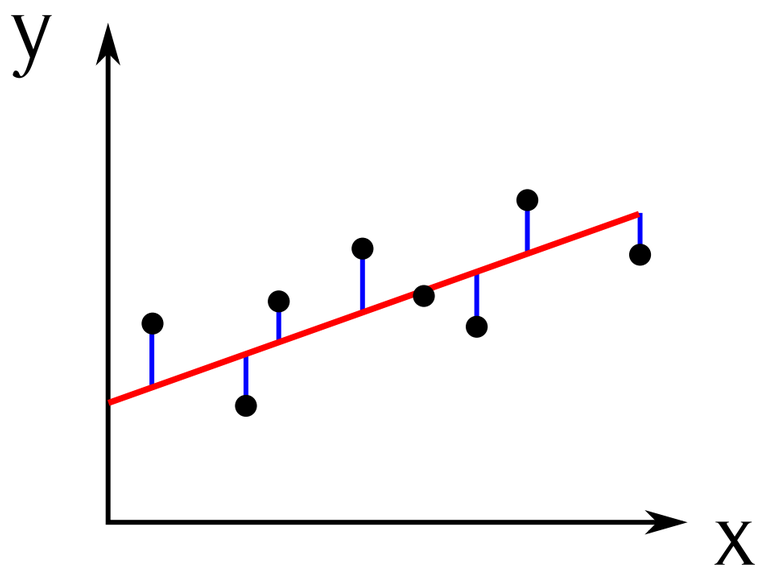
$$y = a + b\ x1 + c\ x2 + d\ x3 + \ldots + \text{erreur}$$

On a des observations de chaque variable

y	x_1	x_2	x_3
30.5	-1.1	-0.1	0.2
16.7	-2.5	-2.4	-2.8
18.3	-3.1	1.1	-1.8
29.1	1.4	1.8	0.8

Quelles sont les valeurs de a, b, c et d les plus *vraisemblables* étant donné ces observations ?

"celles qui minimisent l'écart entre la prédiction et l'observé"



Estimation des paramètres

$$y = a + b\ x1 + c\ x2 + d\ x3 + \ldots + \text{erreur}$$

On a des observations de chaque variable

y	x_1	x_2	x_3
30.5	-1.1	-0.1	0.2
16.7	-2.5	-2.4	-2.8
18.3	-3.1	1.1	-1.8
29.1	1.4	1.8	0.8

Quelles sont les valeurs de a, b, c et d les plus *vraisemblables* étant donné ces observations ?

"celles qui minimisent l'écart entre la prédiction et l'observé"

en pratique...

1. Pour chaque observation i (ligne), on calcule la valeur prédite sans erreur
   $\mu_i = a + b * x1 + c * x2 + d * x3$
2. On additionne tous les carrés des écarts entre y_i et μ_i pour trouver l'écart total
3. On cherche les meilleures valeurs de a, b c et d tels que l'écart total soit minimisé

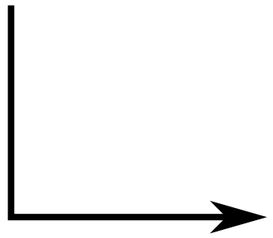
Estimation des paramètres

$$y = a + b\ x1 + c\ x2 + d\ x3 + \ldots + erreur$$

Quel est la probabilité que b soit supérieur à zéro étant donné mes données ?

Quelle est l'erreur estimée sur les paramètres a, b, c et d ?

Si je repète mon expérience, quelles valeurs vais-je vraisemblablement obtenir ?



Estimation par *maximum de vraisemblance*

Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

```
y = a + b x1 + c x2 + d x3 + ... + erreur
```

Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

Pour l'observation (ligne) i:

1/ On calcule la moyenne attendue

$$\mu_i = a + bx_{1,i} + cx_{2,i} + dx_{3,i}$$

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

Pour l'observation (ligne) i:

1/ On calcule la moyenne attendue

$$\mu_i = a + bx_{1,i} + cx_{2,i} + dx_{3,i}$$

2/ On fait l'hypothèse que les observations suivent une loi normale autour de la moyenne:

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

Pour l'observation (ligne) i:

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

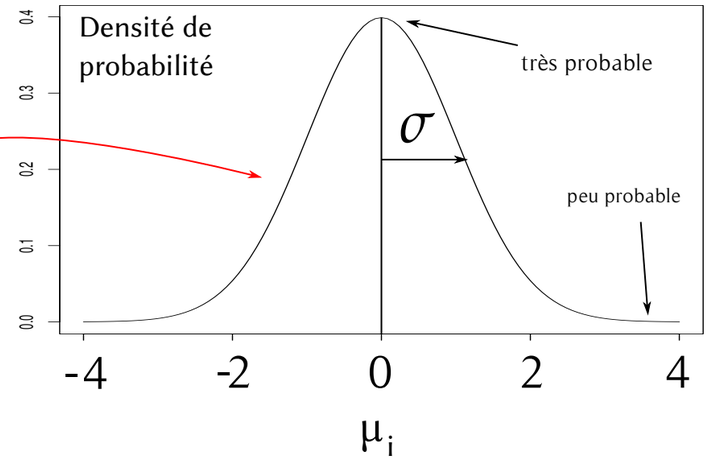
1/ On calcule la moyenne attendue

$$\mu_i = a + bx_{1,i} + cx_{2,i} + dx_{3,i}$$

2/ On fait l'hypothèse que les observations suivent une loi normale autour de la moyenne:

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$P(\mu_i < y_i < \mu_i + \epsilon) = \int_{x=\mu_i}^{x=\mu_i+\epsilon} f(x)dx$$



Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

Pour l'observation (ligne) i:

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

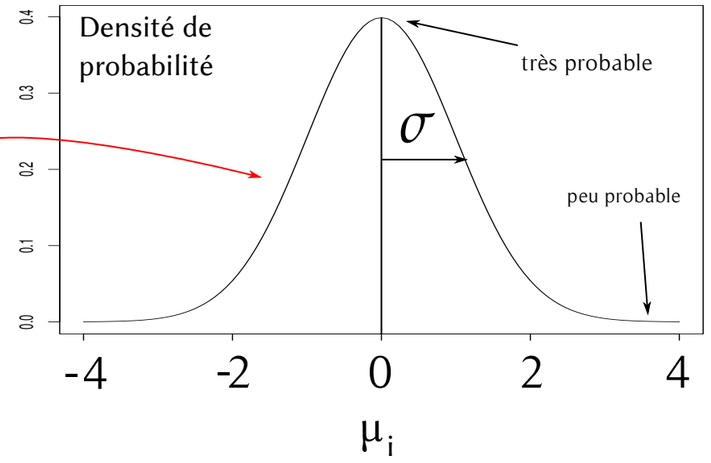
1/ On calcule la moyenne attendue

$$\mu_i = a + bx_{1,i} + cx_{2,i} + dx_{3,i}$$

2/ On fait l'hypothèse que les observations suivent une loi normale autour de la moyenne:

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$P(\mu_i < y_i < \mu_i + \epsilon) = \int_{x=\mu_i}^{x=\mu_i+\epsilon} f(x)dx$$



3/ On calcule la vraisemblance

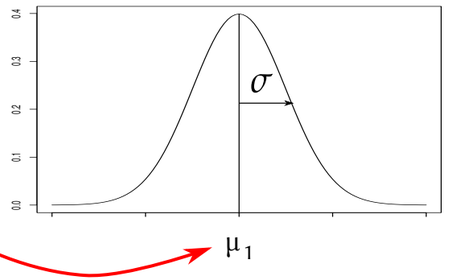
$$\mathcal{L}(a, b, c, d, \sigma) = \prod_{i=1}^{4} P(\mu_i < y_i < \mu_i + \epsilon)$$

Etape 0: On choisit
une valeur numérique pour
a, b, c, d et σ

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

Etape 0: On choisit une valeur numérique pour a, b, c, d et σ

$$\mu_1 = a + b * -1.1 + c * -0.1 + d * 0.2$$

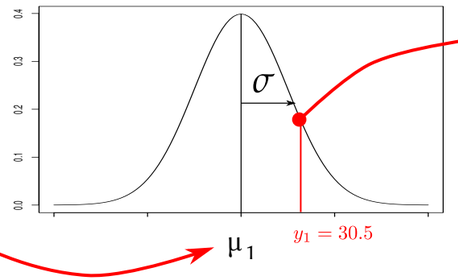


Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

Etape 0: On choisit une valeur numérique pour a, b, c, d et σ

$$\mu_1 = a + b * -1.1 + c * -0.1 + d * 0.2$$

Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8



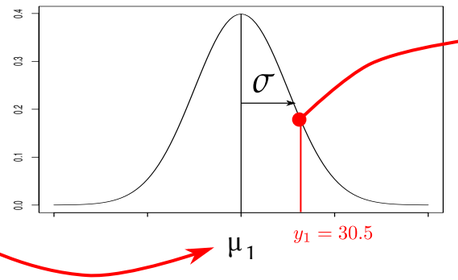
$$P(30.5 < y_1 < 30.5 + \epsilon)$$

Etape 0: On choisit une valeur numérique pour a, b, c, d et σ

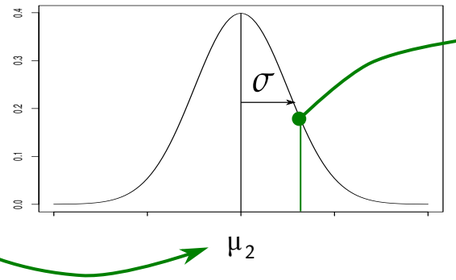
Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

$$\mu_1 = a + b * -1.1 + c * -0.1 + d * 0.2$$

$$\mu_2 = a + b * -2.5 + c * -2.4 + d * -2.8$$



$$P(30.5 < y_1 < 30.5 + \epsilon)$$



$$P(16.7 < y_2 < 16.7 + \epsilon)$$

Etape 0: On choisit
une valeur numérique pour
a, b, c, d et σ

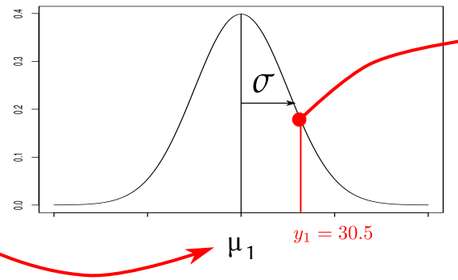
Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

$\mu_1 = a + b * -1.1 + c * -0.1 + d * 0.2$

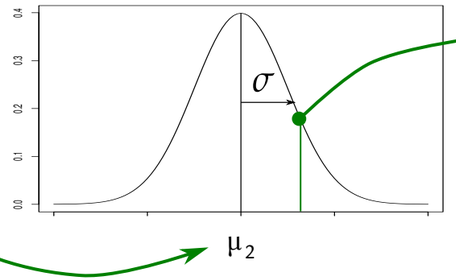
$\mu_2 = a + b * -2.5 + c * -2.4 + d * -2.8$

$\mu_3 = a + b * -3.1 + c * 1.1 + d * -1.8$

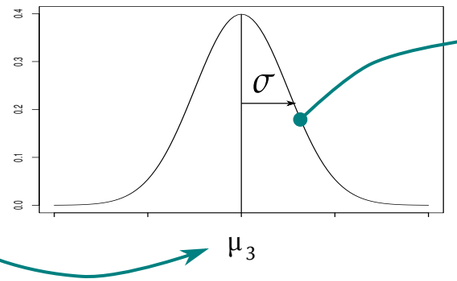
$\mu_4 = a + b * 1.4 + c * 1.8 + d * 0.8$



$P(30.5 < y_1 < 30.5 + \epsilon)$



$P(16.7 < y_2 < 16.7 + \epsilon)$



$P(18.3 < y_3 < 18.3 + \epsilon)$

$P(29.1 < y_4 < 29.1 + \epsilon)$

...

Etape 0: On choisit une valeur numérique pour a, b, c, d et σ

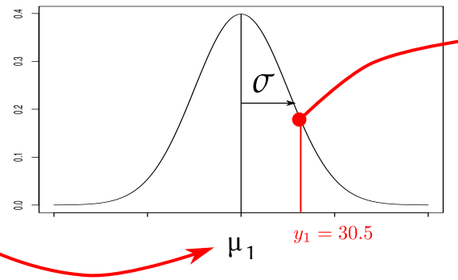
Observation i	y	x_1	x_2	x_3
1	30.5	-1.1	-0.1	0.2
2	16.7	-2.5	-2.4	-2.8
3	18.3	-3.1	1.1	-1.8
4	29.1	1.4	1.8	0.8

$\mu_1 = a + b * -1.1 + c * -0.1 + d * 0.2$

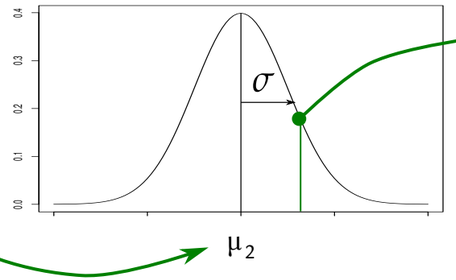
$\mu_2 = a + b * -2.5 + c * -2.4 + d * -2.8$

$\mu_3 = a + b * -3.1 + c * 1.1 + d * -1.8$

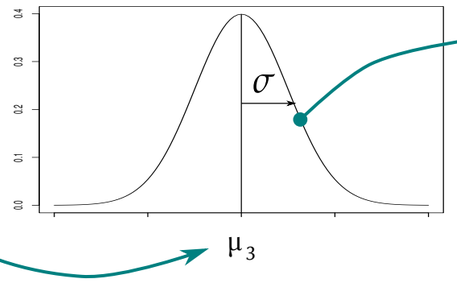
$\mu_4 = a + b * 1.4 + c * 1.8 + d * 0.8$



$P(30.5 < y_1 < 30.5 + \epsilon)$



$P(16.7 < y_2 < 16.7 + \epsilon)$



$P(18.3 < y_3 < 18.3 + \epsilon)$

$P(29.1 < y_4 < 29.1 + \epsilon)$

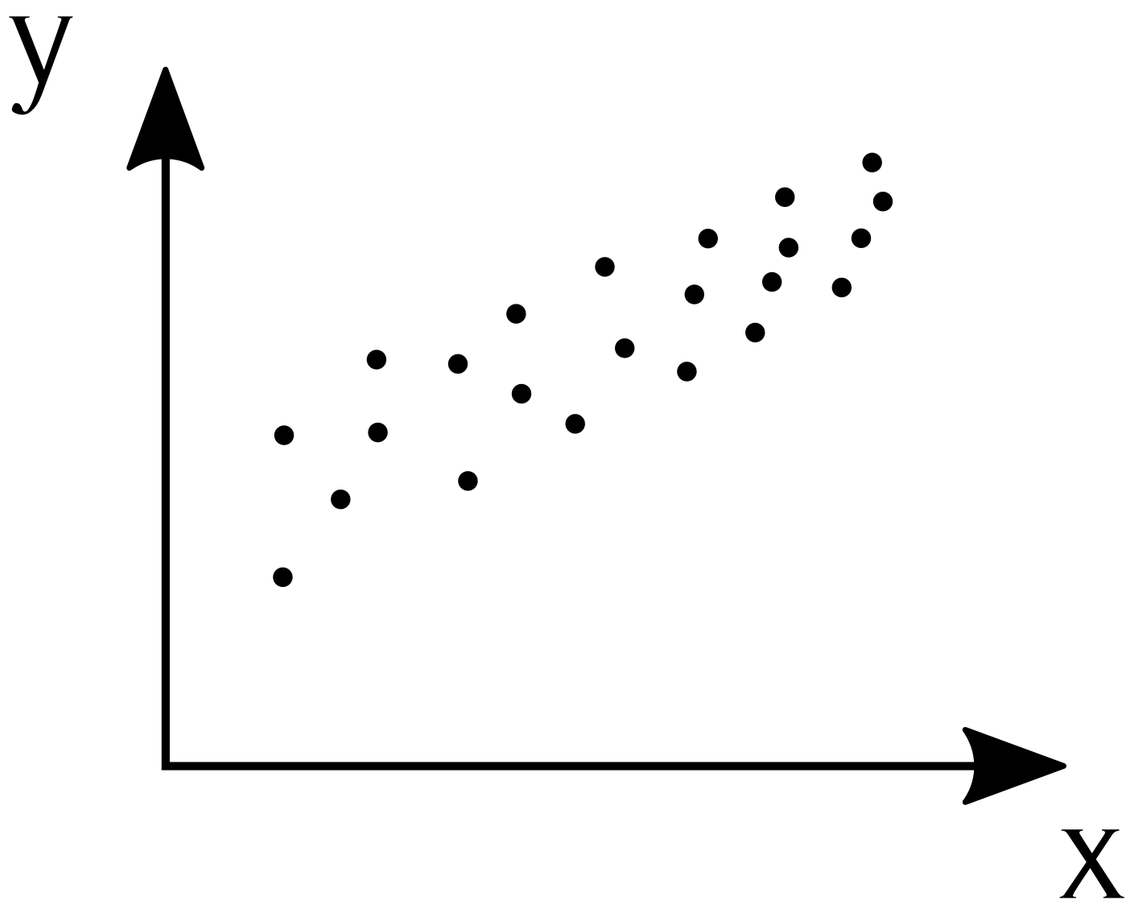
...

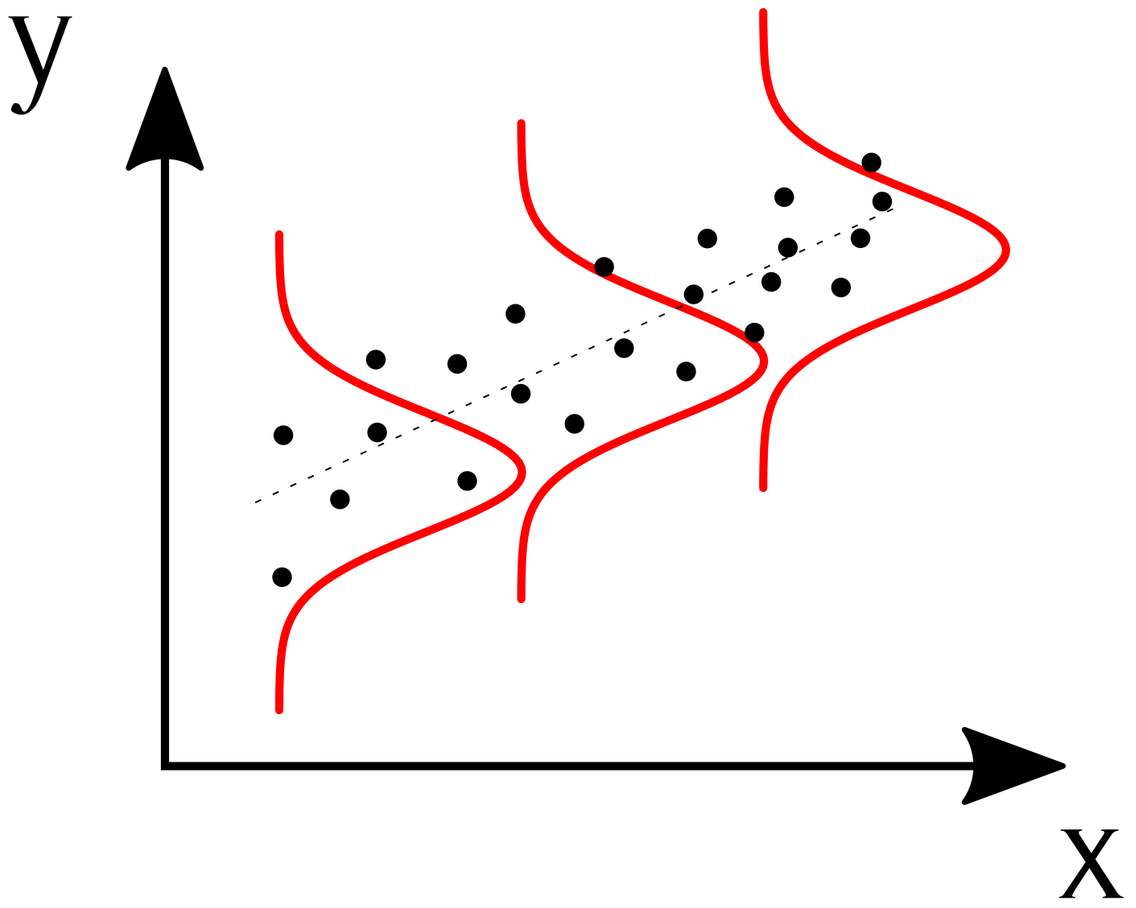
Vraisemblance:

$$\mathcal{L}(a, b, c, d, \sigma) = P(30.5 < y_1 < 30.5 + \epsilon) * P(16.7 < y_2 < 16.7 + \epsilon) * P(18.3 < y_3 < 18.3 + \epsilon) * P(29.1 < y_4 < 29.1 + \epsilon)$$

Estimation par maximum de vraisemblance

La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*



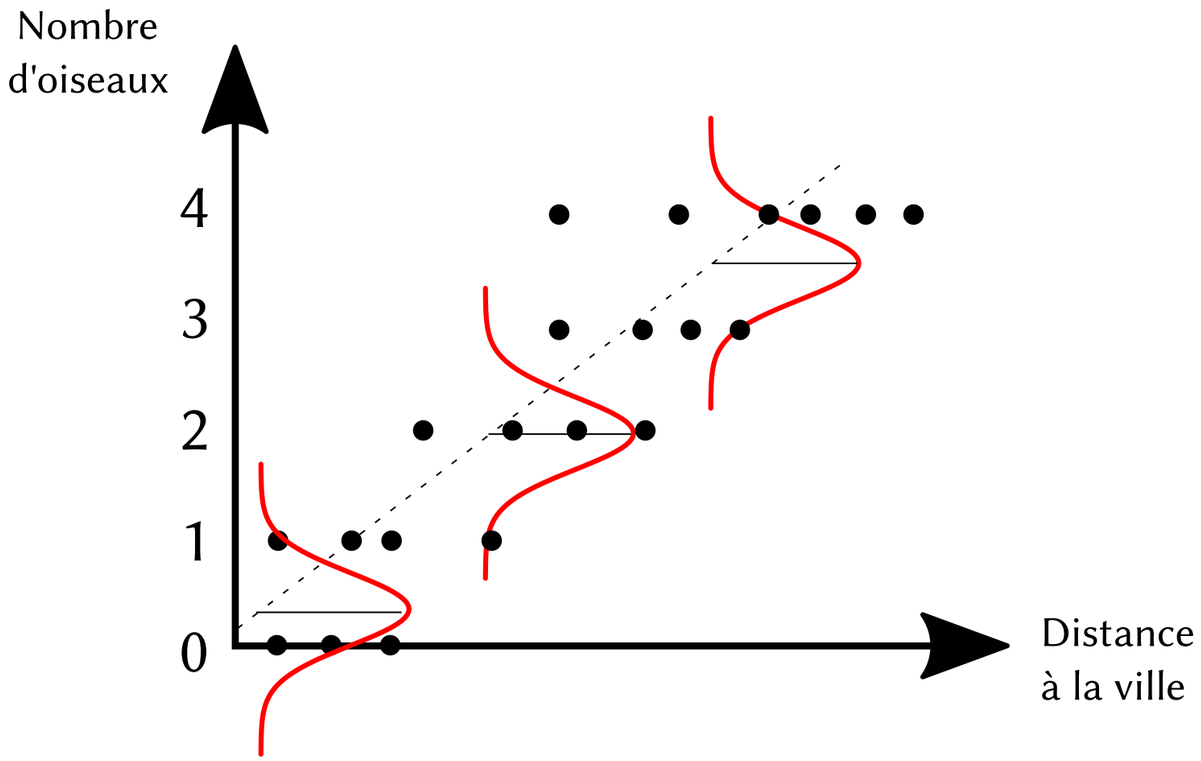


Estimation par maximum de vraisemblance

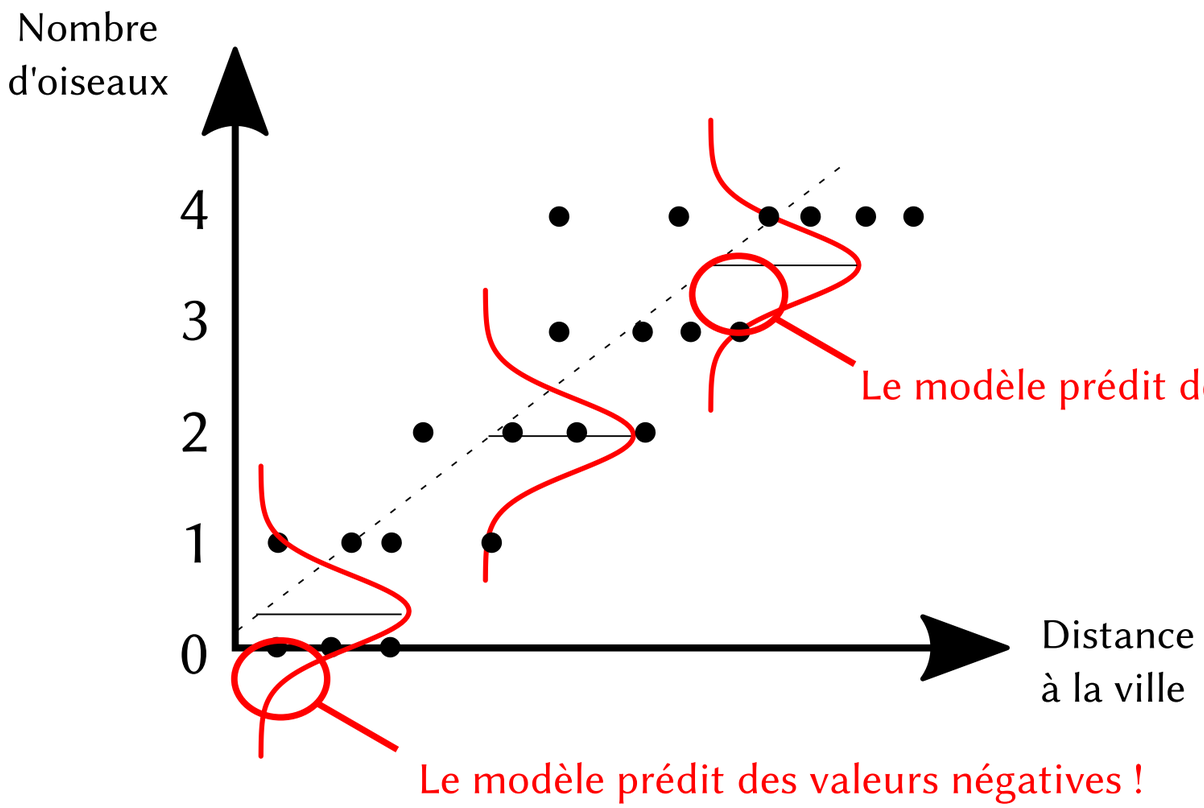
La vraisemblance = *la probabilité d'observer les données étant donnés les paramètres du modèle*

en pratique...

Le problème des modèles linéaires simples

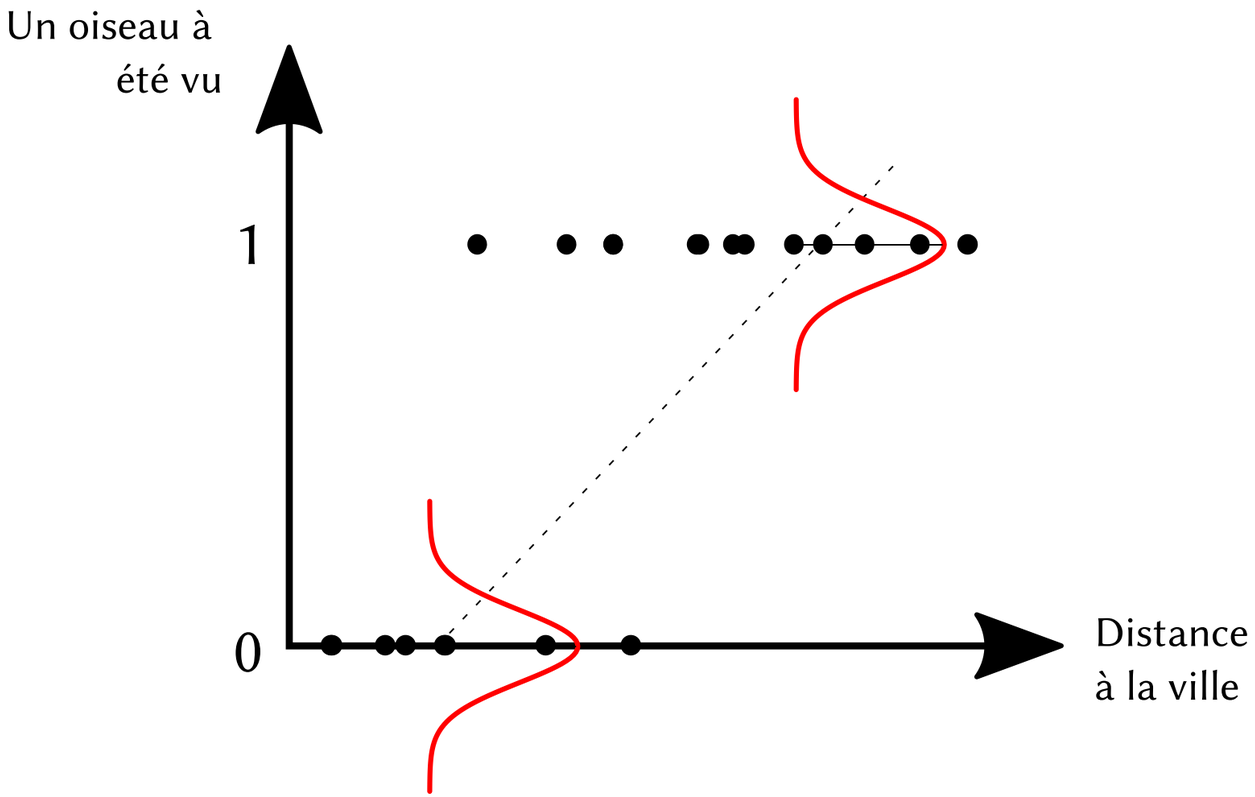


Le problème des modèles linéaires simples



On a besoin d'un modèle
qui prenne en compte la nature
de la variable réponse !

Le problème des modèles linéaires simples



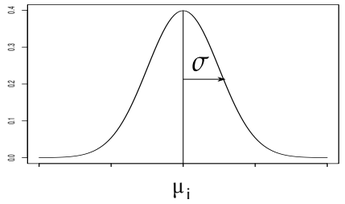
On a besoin d'un modèle qui prenne en compte la nature de la variable réponse !

La solution: le modèle linéaire généralisé ! (*glm*)

Modèle linéaire

$$\mu_i = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Structure d'erreur = loi normale



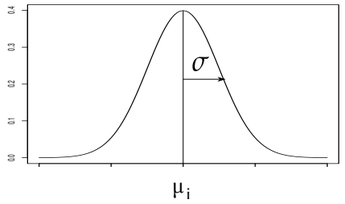
Données continues, entre -∞ et +∞ (ex: température)

La solution: le modèle linéaire généralisé ! (*glm*)

Modèle linéaire

$$\mu_i = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Structure d'erreur = loi normale



Données continues, entre -∞ et +∞ (ex: température)

Modèle linéaire généralisé

"Fonction de lien"

$$g(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

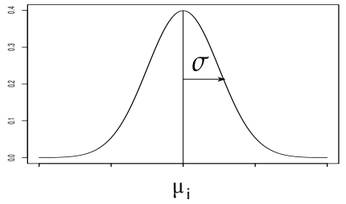
Distribution variable

La solution: le modèle linéaire généralisé ! (*glm*)

Modèle linéaire

$$\mu_i = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Structure d'erreur = loi normale



Données continues, entre -∞ et +∞ (ex: température)

Modèle linéaire généralisé

"Fonction de lien"

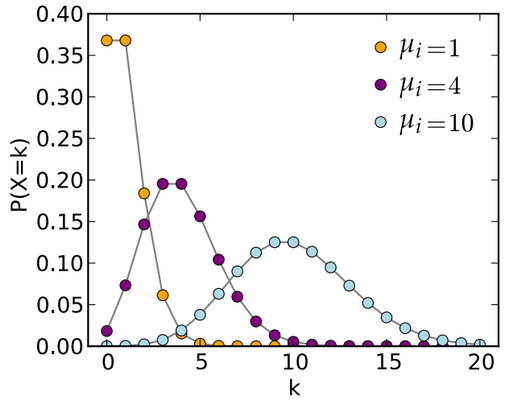
$$g(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

Distribution variable

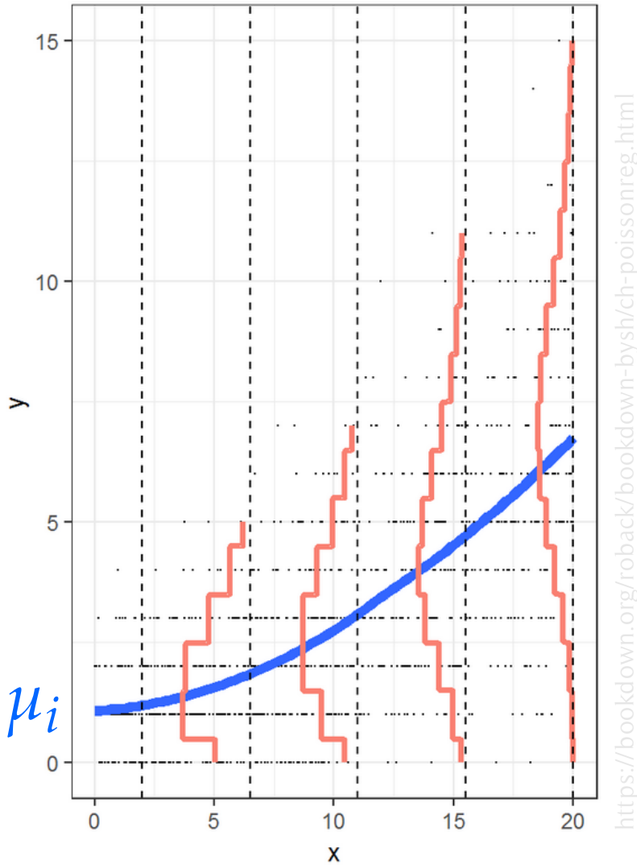
Modèle linéaire généralisé (log-link + poisson)

$$log(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{P}(\mu_i)$$

Structure d'erreur = loi de poisson



Données discrètes, entre 0 et +∞ (ex: comptages)



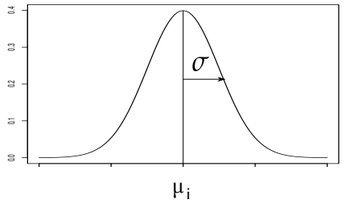
https://bookdown.org/roback/bookdown-bysh/ch-poissonreg.html

La solution: le modèle linéaire généralisé ! (*glm*)

Modèle linéaire

$$\mu_i = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Structure d'erreur =
loi normale



Données continues,
entre -∞ et +∞
(ex: température)

Modèle linéaire généralisé

"Fonction de lien"

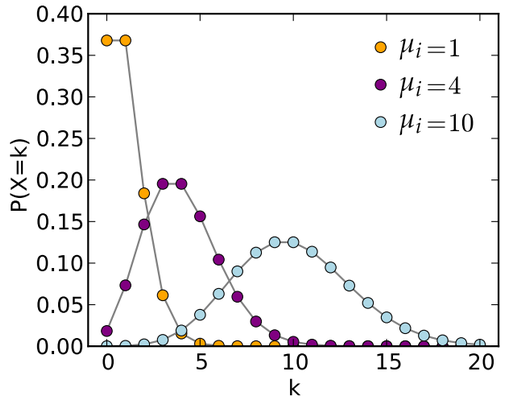
$$g(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

Distribution variable

Modèle linéaire généralisé
(log-link + poisson)

$$log(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{P}(\mu_i)$$

Structure d'erreur =
loi de poisson



Données discrètes,
entre 0 et +∞
(ex: comptages)

Modèle linéaire généralisé
(logit-link + bernouilli distribution)

$$logit(\mu_i) = a + bx_{1,i} + cx_{2,i}$$
$$y_i \sim \mathcal{B}(\mu_i)$$

Structure d'erreur =
loi de bernouilli

1 avec un probabilité μ_i
0 avec un probabilité 1 - μ_i

Données catégoriques
à deux catégories (0/1)
(ex: succès/échec)

extensible au
nombre de succès parmi N

La solution: le modèle linéaire généralisé ! (*glm*)

en pratique...

La solution: le modèle linéaire généralisé ! (*glm*)

Les variables catégoriques ?

Observation i	y	x_1	c_1
1	30.5	-1.1	grand
2	16.7	-2.5	petit
3	18.3	-3.1	petit
4	29.1	1.4	grand

$$g(\mu_i) = a + bx_{1,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

2 options

La solution: le modèle linéaire généralisé ! (*glm*)

Les variables catégoriques ?

$$g(\mu_i) = a + bx_{1,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

Observation i	y	x_1	c_1
1	30.5	-1.1	grand
2	16.7	-2.5	petit
3	18.3	-3.1	petit
4	29.1	1.4	grand

$$g(\mu_i) = a + bx_{1,i} + d_{[c_1=grand]}$$

ou

$$g(\mu_i) = a + bx_{1,i} + d_{[c_1=petit]}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

l'effet b de la variable x_1 n'est pas affectée par c_1,
seul l'*intercept* a l'est

Sous R: `y ~ 1 + x1 + c1`

La solution: le modèle linéaire généralisé ! (*glm*)

Les variables catégoriques ?

Observation i	y	x_1	c_1
1	30.5	-1.1	grand
2	16.7	-2.5	petit
3	18.3	-3.1	petit
4	29.1	1.4	grand

$$g(\mu_i) = a + bx_{1,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

2 options

$$g(\mu_i) = a + bx_{1,i} + d_{[c_1=grand]}$$
ou
$$g(\mu_i) = a + bx_{1,i} + d_{[c_1=petit]}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

l'effet b de la variable x_1 n'est pas affectée par c_1,
seul l'*intercept* a l'est

Sous R: `y ~ 1 + x1 + c1`

$$g(\mu_i) = a + b_{[c1=petit]}x_{1,i}$$
ou
$$g(\mu_i) = a + b_{[c1=grand]}x_{1,i}$$
$$y_i \sim \mathcal{D}(\mu_i, ...)$$

l'effet b de la variable x_1 est affectée par c_1,
= une *interaction*

`y ~ 1 + x1:c1` (pente seulement)
`y ~ 1 + c1 + x1:c1` (pente + intercept)

La solution: le modèle linéaire généralisé ! (*glm*)

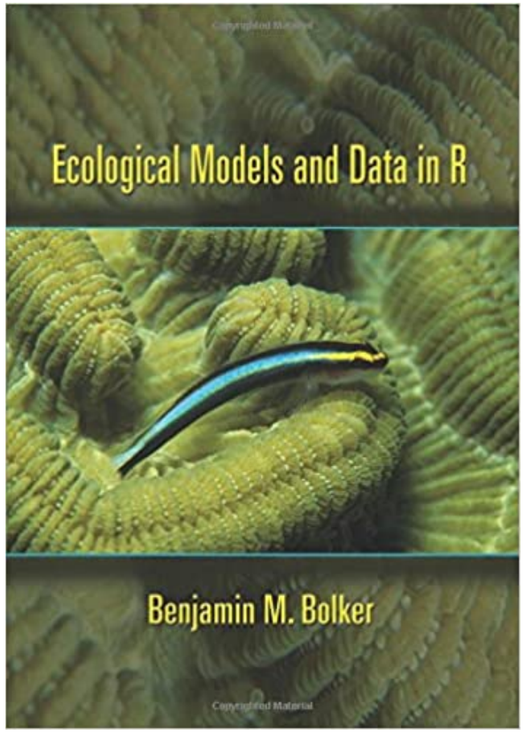
en pratique...

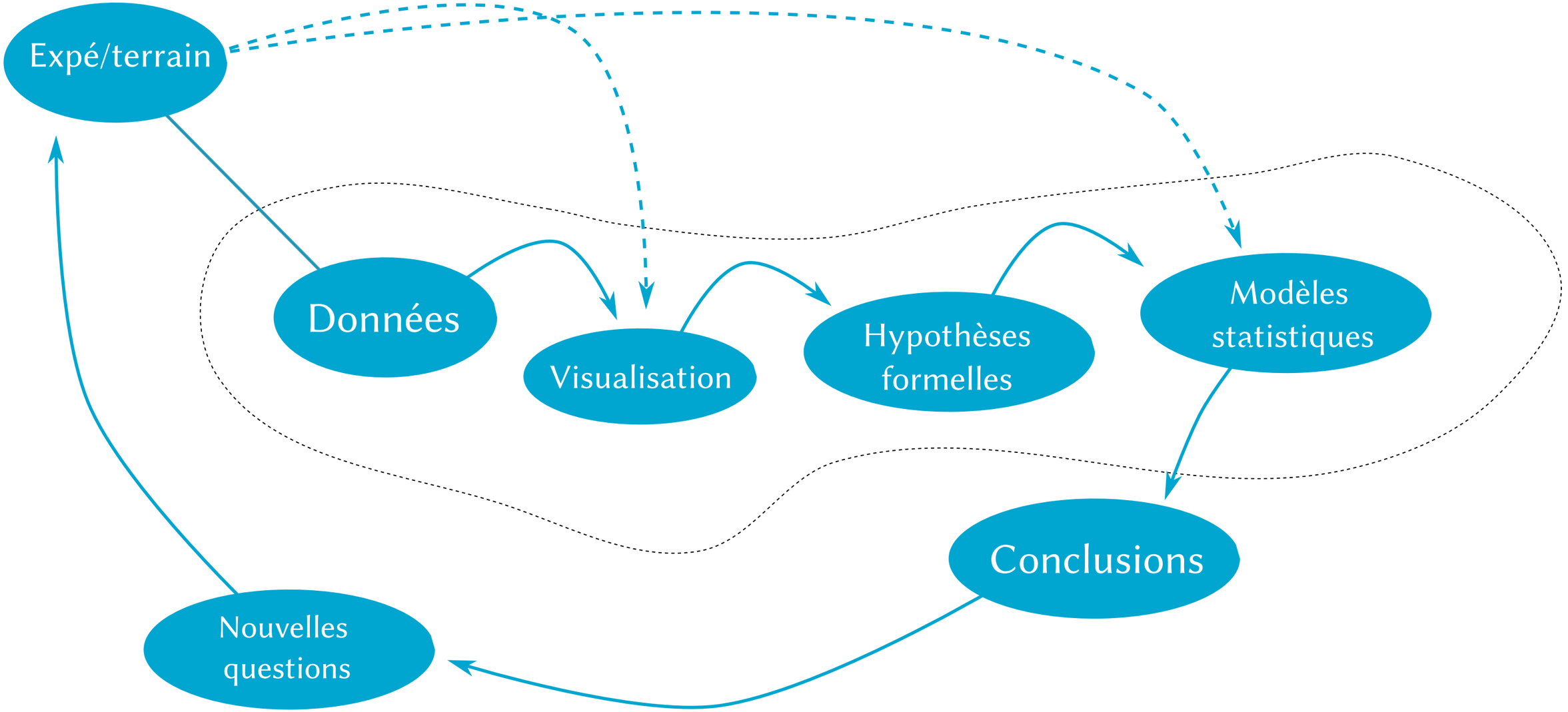
Le monde des modèles linéaires est vaste

Visualiser et connaître son objet d'étude !

Lire et prendre des cours !

Que font les autres dans mon domaine ?




Expé/terrain
Données
Visualisation
Hypothèses formelles
Modèles statistiques
Conclusions
Nouvelles questions

Recap:

- **Aperçu rapide des modèles linéaires (généralisés) dans R**
- **Implémentation et lecture des résultats**
- **Pas la sélection de modèles**
- **Pas les modèles mixtes**

Atelier: jeudi 19 novembre, 14h !

double atelier !

Tous les exercices et infos sur https://rrr.mbb.cnrs.fr