La régression linéaire simple

Modélisation par une relation linéaire de l'évolution d'une variable quantitative observée en fonction d'une variable quantitative contrôlée

M. L. Delignette-Muller
VetAgro Sup

16 décembre 2019

Objectifs pédagogiques

- Connaître le modèle utilisé en régression linéaire simple et la méthode d'estimation de ses paramètres à partir de données.
- Savoir expliquer ce que représente la valeur de r^2.
- Savoir identifier les cas sur lesquels il convient d'utiliser une régression linéaire et dans ces cas distinguer la variable indépendante et la variable dépendante.
- Savoir interpréter les résultats d'une régression linéaire issus d'un logiciel et vérifier ses conditions d'utilisation.
- Savoir utiliser un modèle de régression linéaire en prédiction (avec distinction entre les deux intervalles de confiance).
- Ne pas confondre régression et corrélation linéaire.
- Avoir un aperçu du champ d'utilisation du modèle linéaire et de ses extensions.

Plan

Exemple inspiré de la littérature

Roomi *et al.* 2011, Nutrient mixture inhibits *in vitro* and *in vivo* growth of human acute promyelocytic leukemia HL-60 cells, *Experimental Oncology*
Impact, *in vitro*, d'un mélange de nutriments (acide ascorbique, extrait de thé vert, lysine, proline, ...) sur la prolifération de cellules tumorales.

- **Variable contrôlée notée** X : dose de nutriments en concentration dans le milieu ($\mu g.ml^{-1}$)
- **Variable observée notée** Y : prolifération cellulaire quantifiée en pourcentage de celle observée sans nutriments dans le milieu de culture

Les données de l'expérience réalisée

Données brutes (telles que saisies informatiquement) :

dose	proliferation
10	105
10	92
10	103
50	85
50	89
50	76
100	60
100	72
100	80
500	55
500	48
500	41
1000	42
1000	38
1000	21

...

Diagramme de dispersion (ou nuage de points)

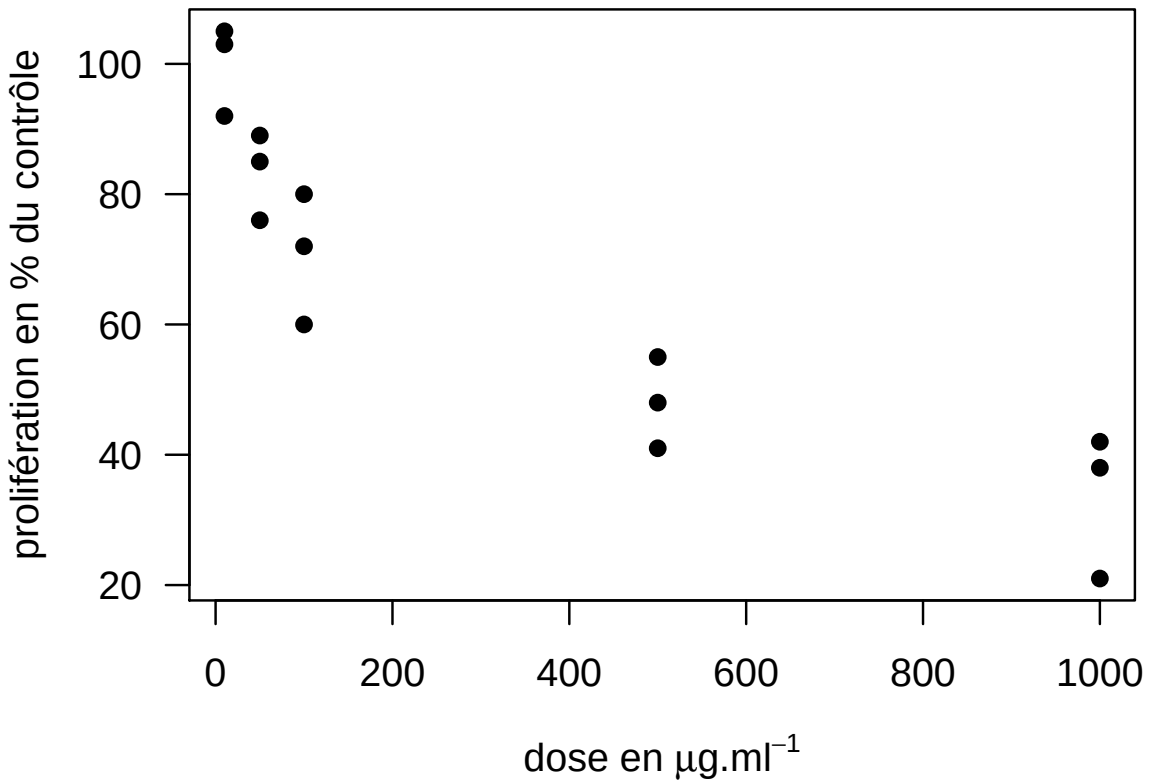
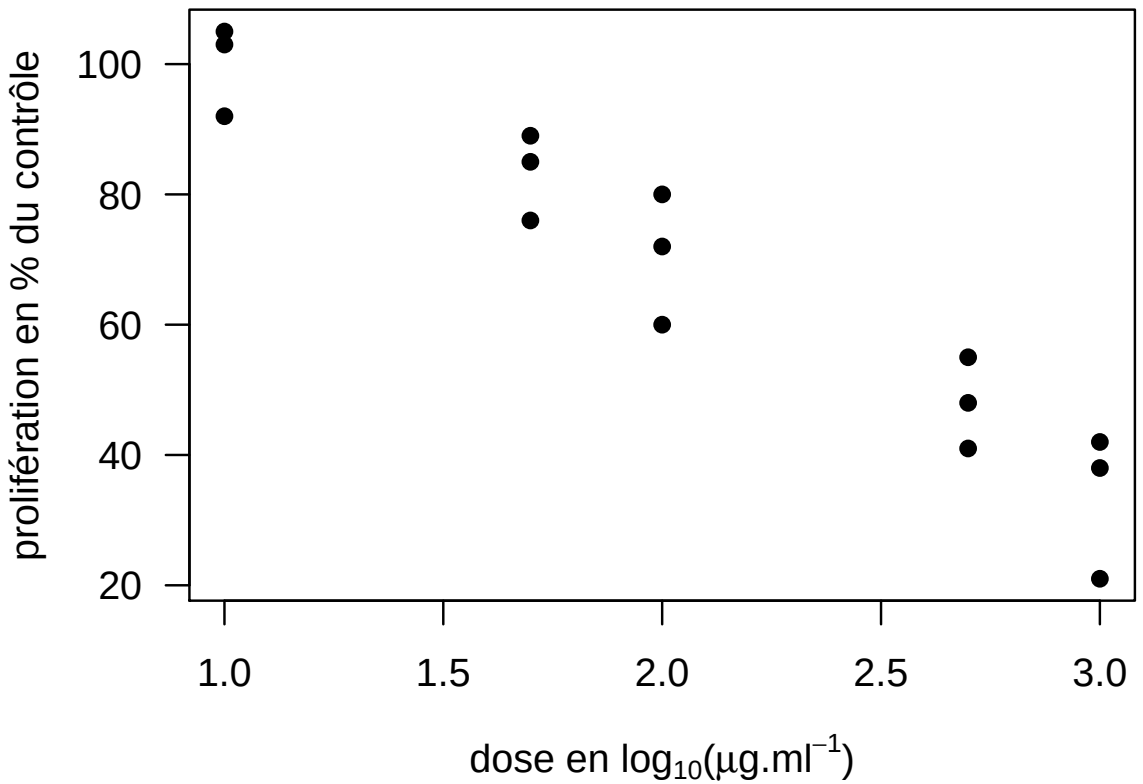


Diagramme de dispersion (ou nuage de points) des données transformées

Après transformation logarithmique de la variable de contrôle dans cet exemple pour linéariser la relation.

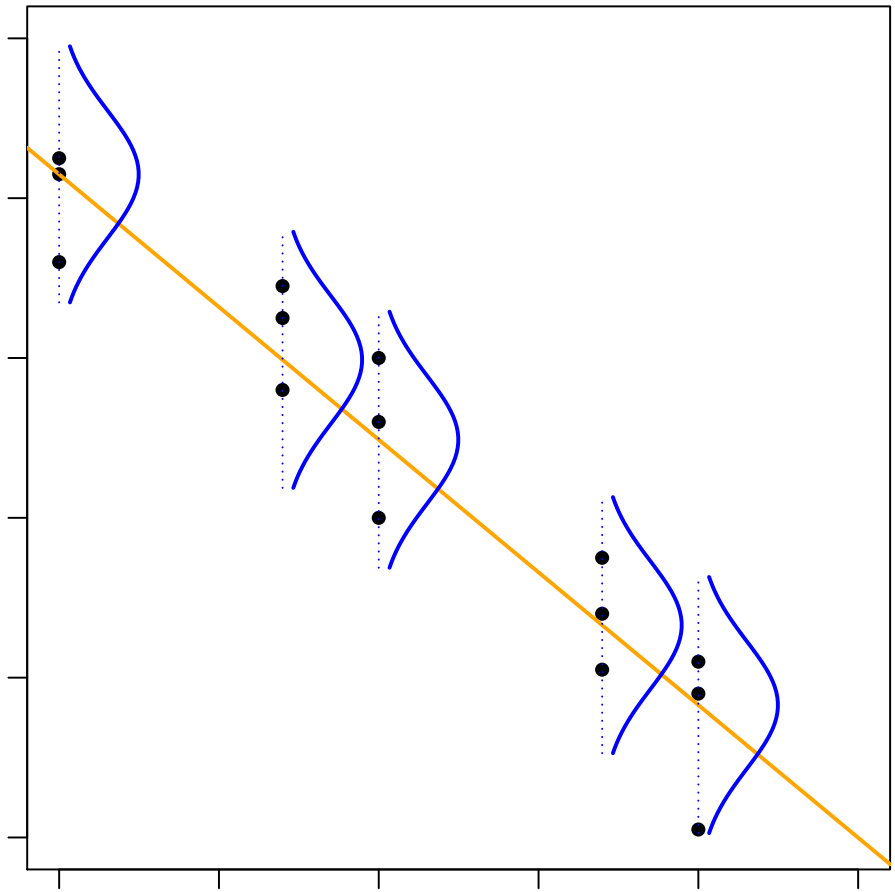


Variable indépendante X et variable dépendante Y

Plus généralement en régression linéaire simple on utilise un modèle linéaire pour expliquer
une variable observée notée Y, appelée aussi variable **à expliquer** ou variable **dépendante**
en fonction
d'une variable explicative notée X (souvent contrôlée mais pas toujours), appelée aussi variable **variable indépendante**.

Le modèle théorique

$Y_i = \alpha + \beta X_i + \epsilon_i$
avec $\epsilon_i \sim N(0, \sigma)$
Partie déterministe : relation linéaire
Partie stochastique : modèle gaussien
ϵ_i aléatoires, indépendants, suivant une loi normale (loi de Gauss) de variance résiduelle σ^2 constante.



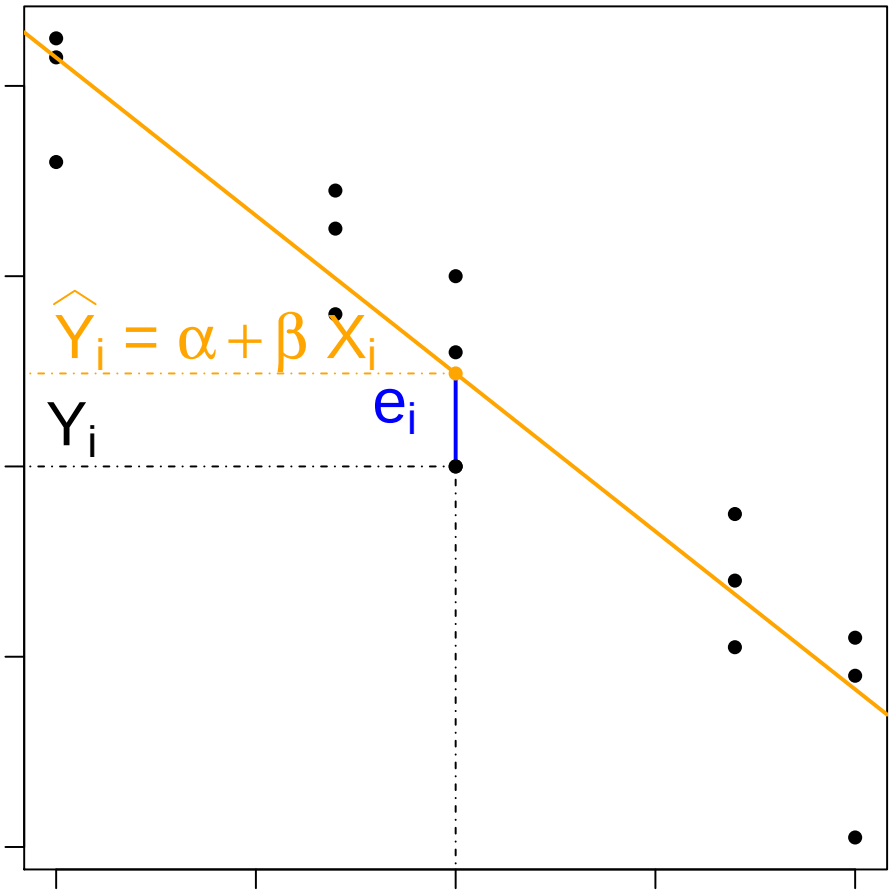
Méthode d'estimation des paramètres

Maximisation de la vraisemblance ($Pr(Y|\alpha, \beta, \sigma)$) qui revient dans le cadre du modèle gaussien à la minimisation de la Somme des Carrés des Ecarts (SCE)

$SCE = \sum_{i=1}^{n} e_i^2$

avec

$e_i = Y_i - \hat{Y}_i$



Estimation ponctuelle des paramètres

- Pente (ou coefficient de régression) :
  $\hat{\beta} = \frac{cov(X,Y)}{V(X)}$
- Ordonnée à l'origine ("intercept" en anglais) :
  $\hat{\alpha} = \overline{Y} - \hat{\beta} \times \overline{X}$
- Ecart type résiduel ("residual standard error" en anglais) :
  $\hat{\sigma} = \sqrt{\frac{SCE}{n-2}}$

Estimation des paramètres avec R

```
> m <- lm(proliferation ~ log10(dose), data = d)
> summary(m)

Call:
lm(formula = proliferation ~ log10(dose), data = d)

Residuals:
   Min     1Q Median     3Q    Max
-15.57  -4.67   1.43   5.33  10.22

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    136.18       6.37    21.4  1.6e-11
log10(dose)    -33.20       2.90   -11.5  3.6e-08

Residual standard error: 8.01 on 13 degrees of freedom
Multiple R-squared:  0.91,        Adjusted R-squared:  0.903
F-statistic:  131 on 1 and 13 DF,  p-value: 3.64e-08
```

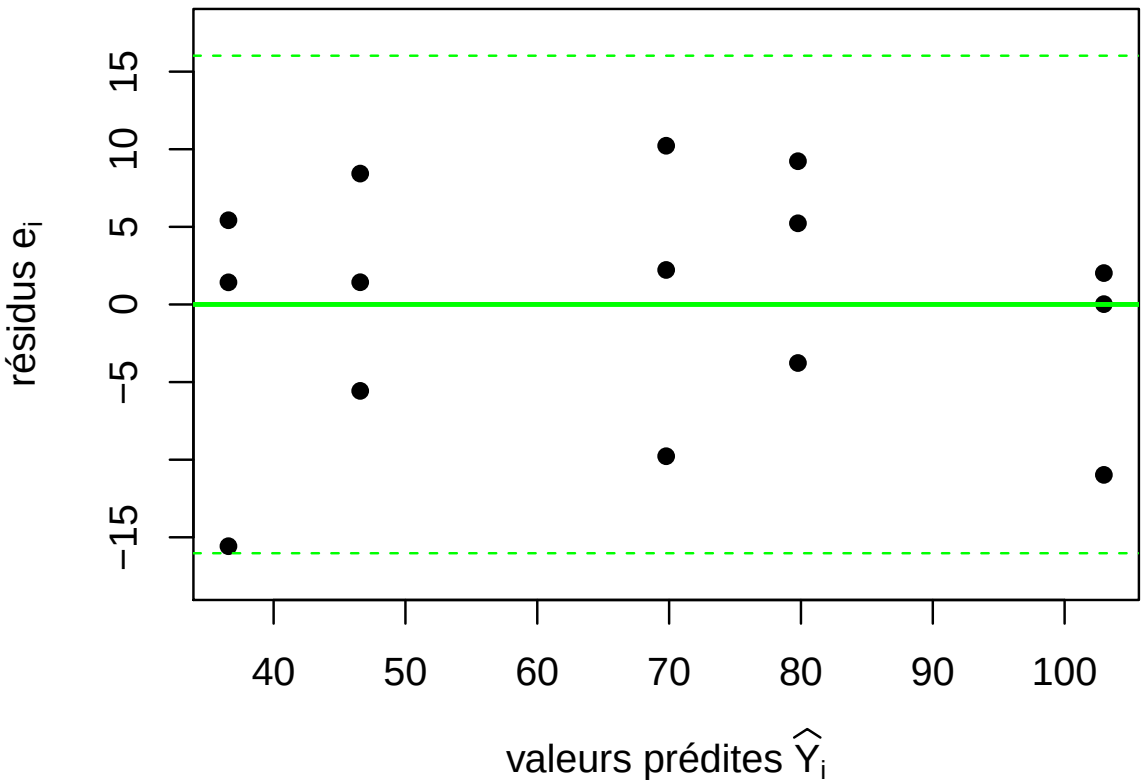
Estimation des paramètres avec **R** - lecture des éléments principaux du résumé fourni par **R** sur cet exemple

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    136.18       6.37    21.4  1.6e-11
log10(dose)    -33.20       2.90   -11.5  3.6e-08

Residual standard error: 8.01 on 13 degrees of freedom
Multiple R-squared:  0.91,        Adjusted R-squared:  0.903
F-statistic:  131 on 1 and 13 DF,  p-value: 3.64e-08
```

- pente : $\hat{\beta} = -33.2$
- ordonnée à l'origine : $\hat{\alpha} = 136$
- écart type résiduel : $\hat{\sigma} = 8.01$
- et le coefficient de détermination (carré du coefficient de corrélation, dont on expliquera le sens plus loin) : $r^2 = 0.91$

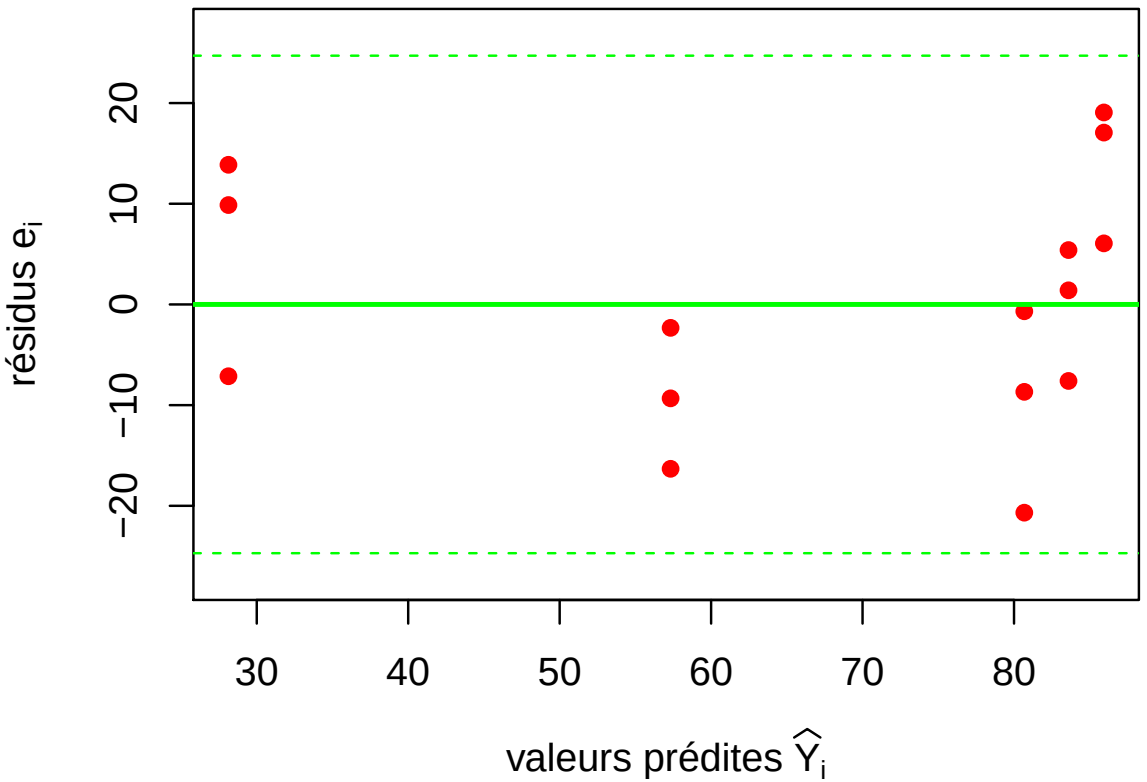
Vérification *a posteriori* des conditions d'utilisation

On s'attend à une répartition aléatoire des résidus selon une loi normale de variance σ^2 constante (environ 95% des résidus dans $[-2\sigma; 2\sigma]$).



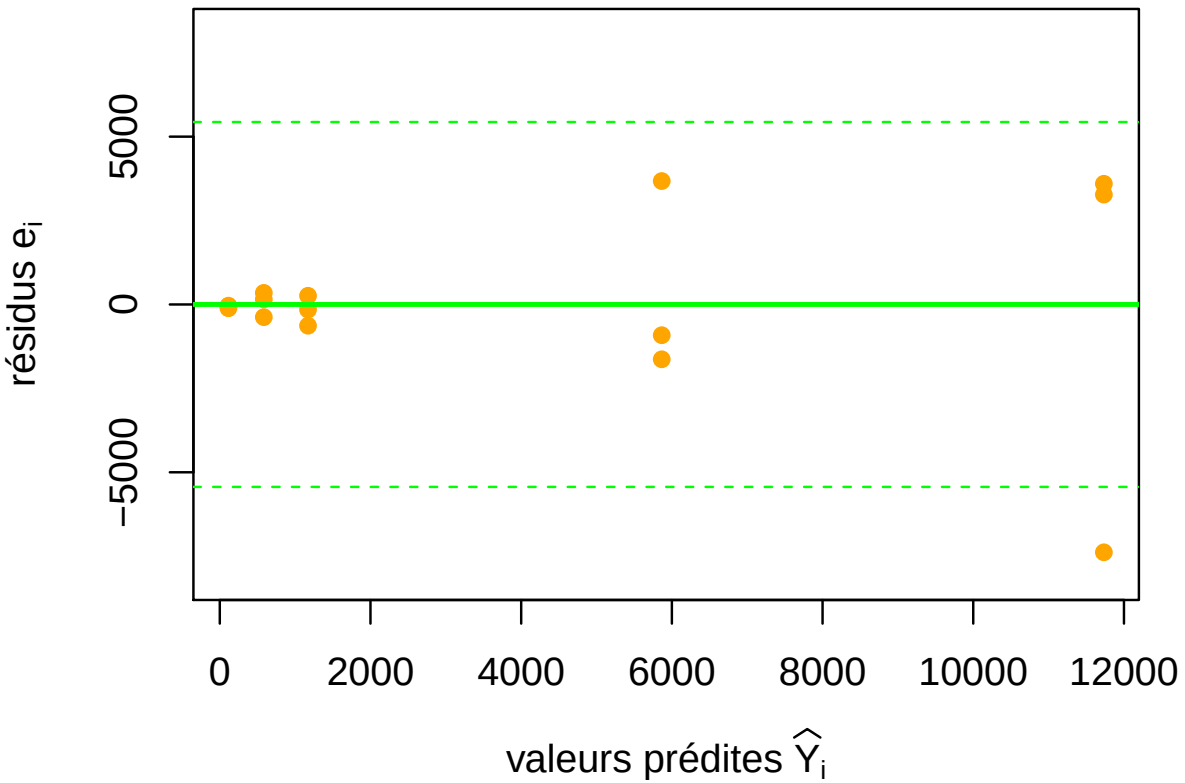
Exemple b de mauvais graphe des résidus

Exemple de régression sans tranformation logarithmique des doses : graphe des résidus amenant à rejeter le modèle du fait du caractère non aléatoire des résidus



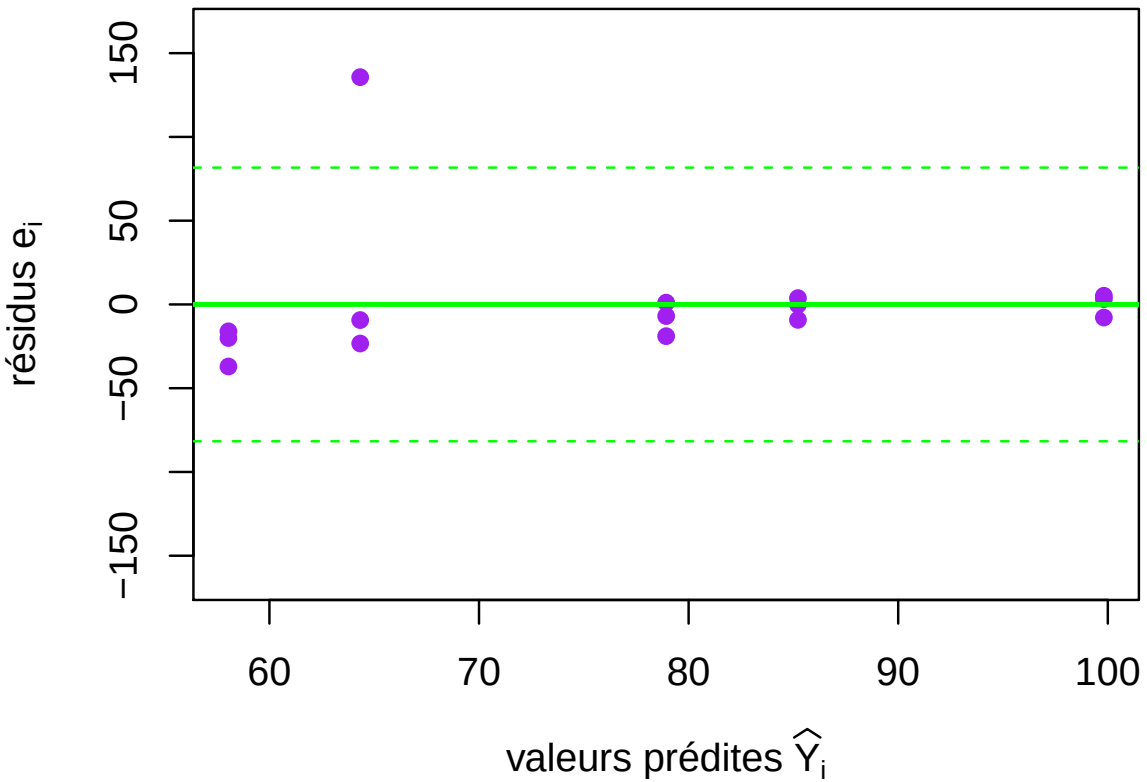
Exemple c de mauvais graphe des résidus

Régression sur un jeu de données différent : graphe des résidus amenant à rejeter le modèle du caractère non constant de la variance résiduelle (hétéroscédasticité)



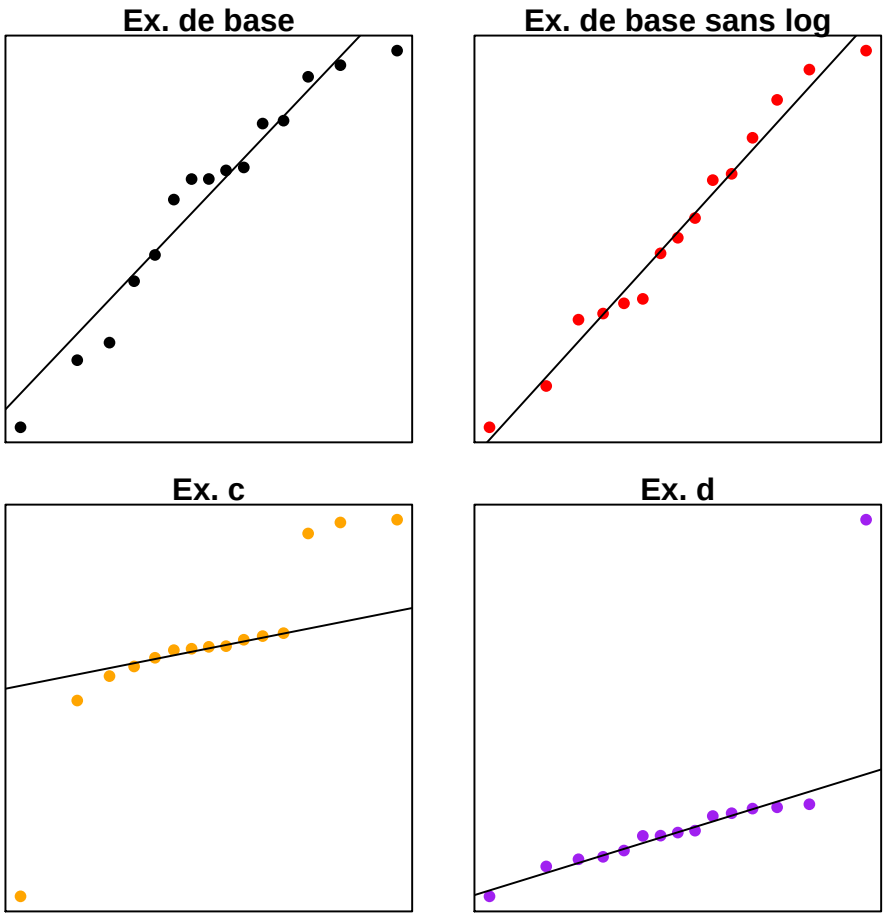
Exemple d de mauvais graphe des résidus

Régression sur un jeu de données différent : graphe des résidus amenant à remettre en question l'ajustement du modèle aux données du fait d'une valeur extrême



Diagrame Quantile - Quantile des résidus

Graphe complémentaire sur lequel on attend des points alignés.



Ce graphe complémentaire sert à vérifier l'hypothèse de normalité de l'ensemble des résidus (remise en cause ici sur les Ex. c et d) mais ne permet pas forcément de détecter un problème lié au caractère non aléatoire des résidus (Ex. de base sans log).

Plan

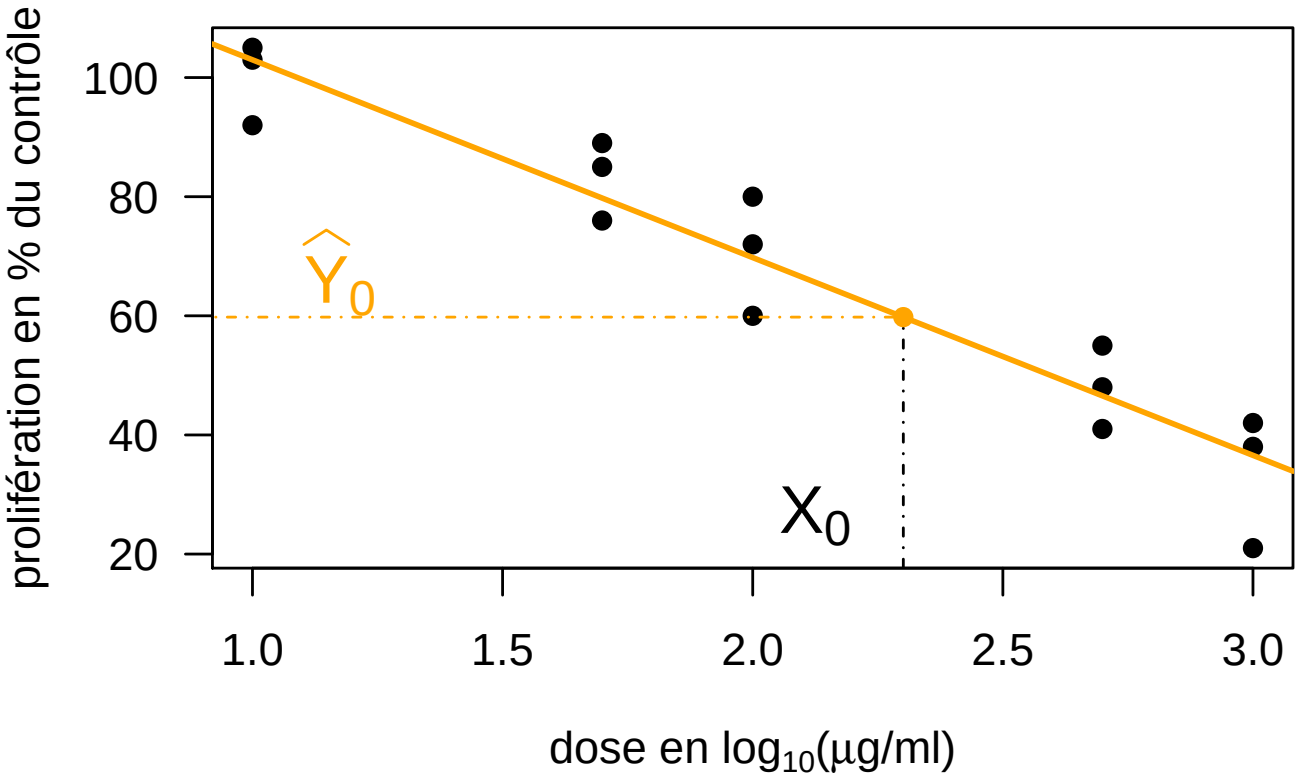
Estimation par intervalle des paramètres du modèle

Si les paramètres du modèle sont utilisés directement, il est important d'associer à leur estimation ponctuelle un intervalle de confiance.
Dans R :

```
> (m <- lm(proliferation ~ log10(dose), data = d))

Call:
lm(formula = proliferation ~ log10(dose), data = d)

Coefficients:
(Intercept)  log10(dose)
      136.2        -33.2

> confint(m)

             2.5 % 97.5 %
(Intercept) 122.4  149.9
log10(dose) -39.5  -26.9
```

Prédiction à partir du modèle

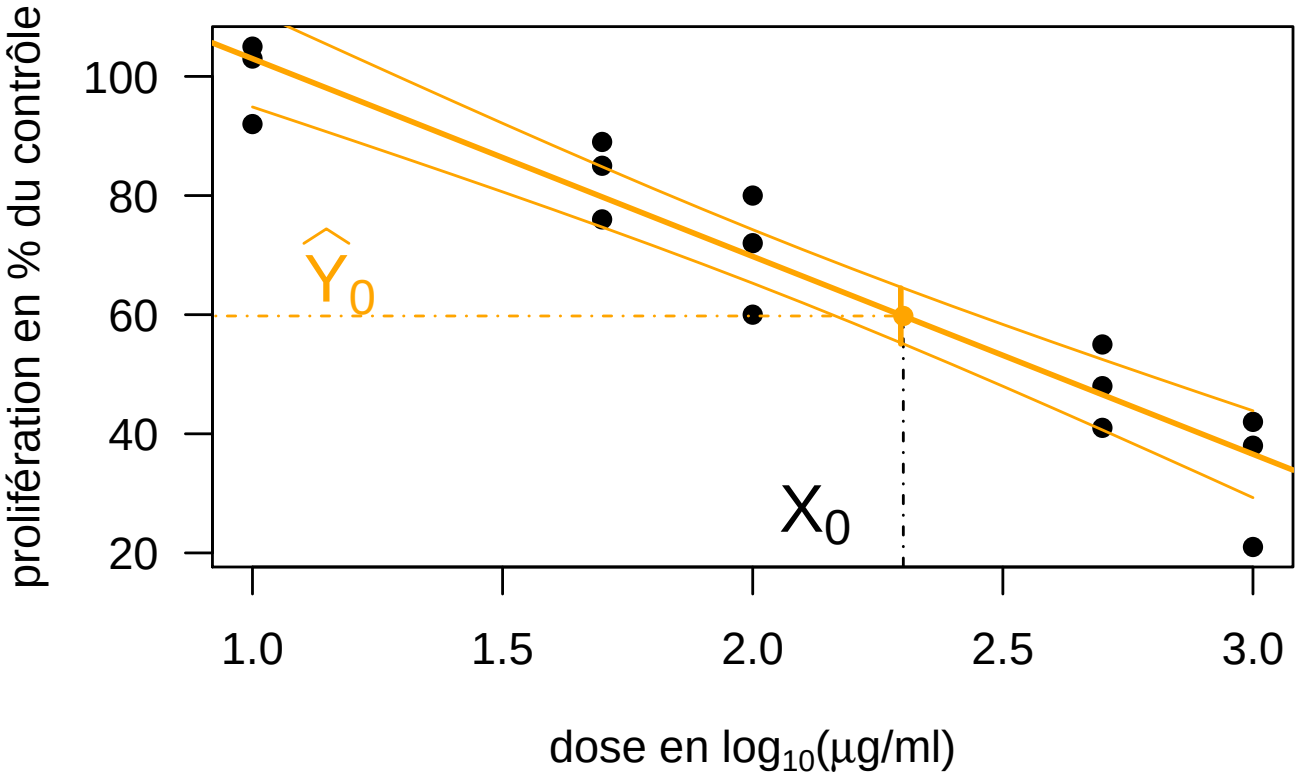
Prédiction d'une valeur de Y_0 pour $X = X_0$ dans le domaine étudié.



Prédiction à partir du modèle - intervalle de confiance

Prédiction d'une valeur de Y_0 pour $X = X_0$ dans le domaine étudié.

Intervalle de confiance sur la moyenne (incertitude sur la droite)

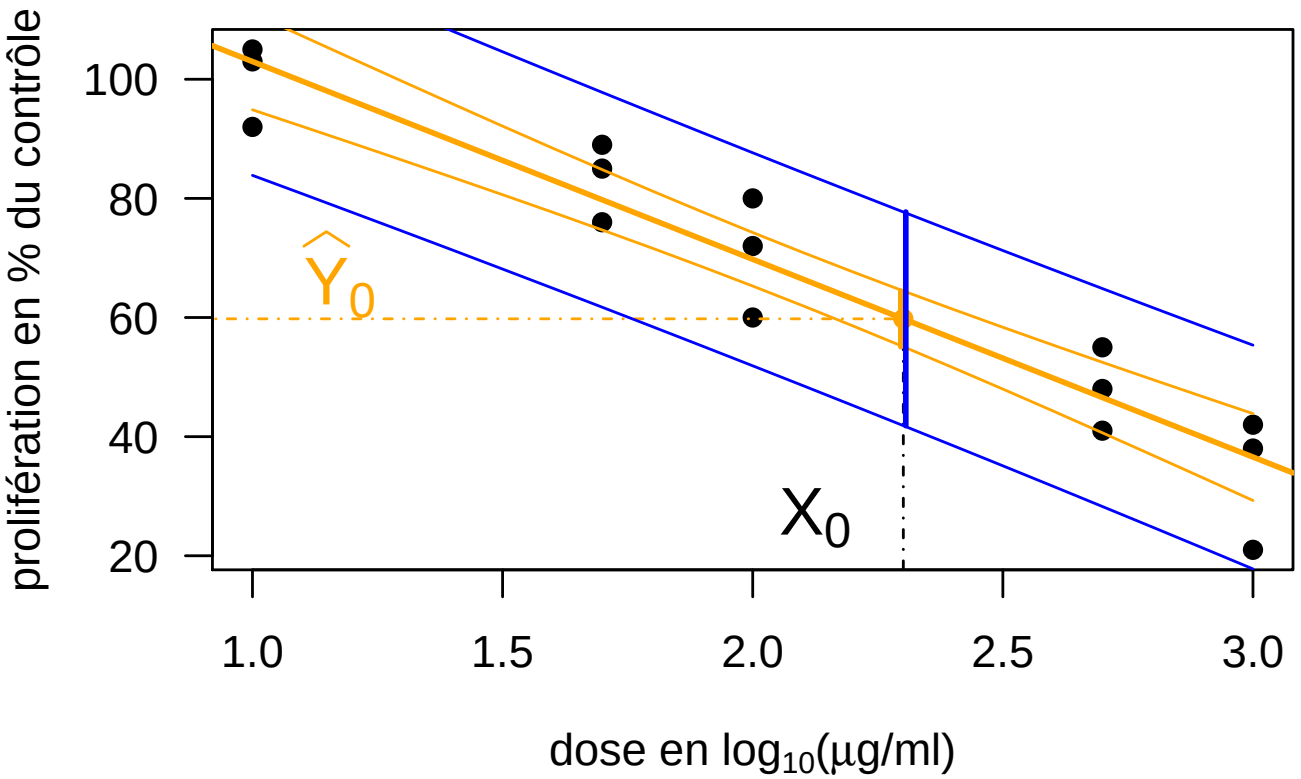


Prédiction à partir du modèle - intervalle de prédiction

Prédiction d'une valeur de Y_0 pour $X = X_0$ dans le domaine étudié.

Intervalle de confiance sur la moyenne (marge d'erreur sur la droite)

Intervalle de prédiction (marge d'erreur sur une observation prédite)



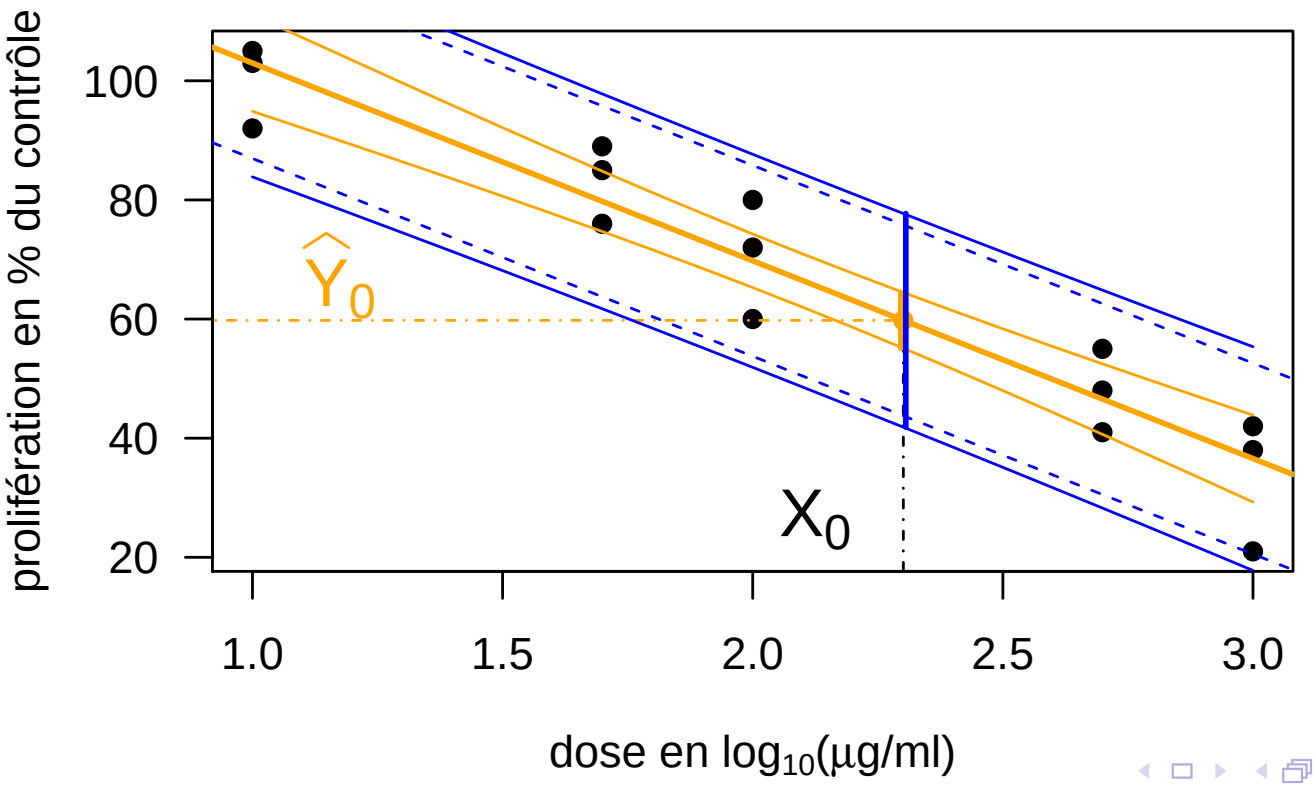
Prédiction à partir du modèle - intervalle de prédiction

Prédiction d'une valeur de Y_0 pour $X = X_0$ dans le domaine étudié.

Intervalle de confiance sur la moyenne (marge d'erreur sur la droite)

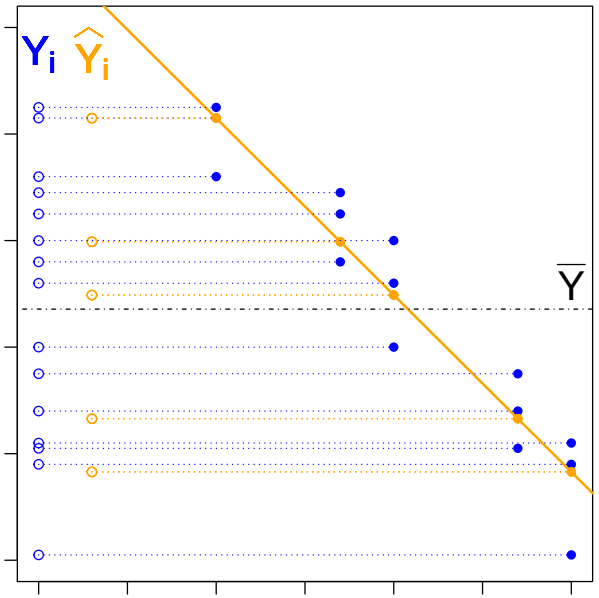
Intervalle de prédiction (marge d'erreur sur une observation prédite)

approché souvent par $\hat{Y}_0 \pm 2 \times \sigma$ (en pointillés)



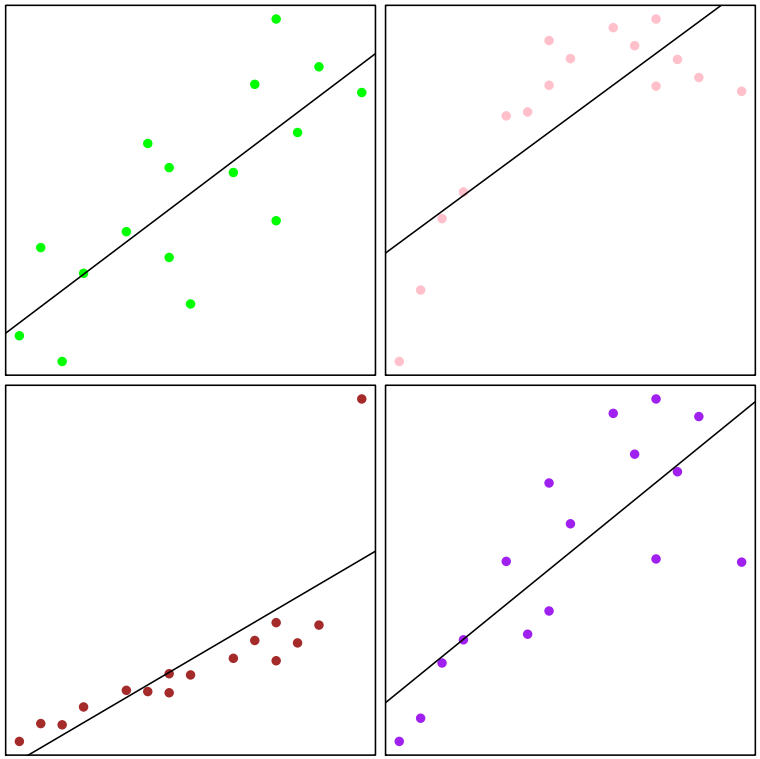
Coefficient de détermination : r^2 part de variation expliquée par le modèle

Soit $r = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}}$ le coefficient de corrélation linéaire, r^2 est le rapport de la variation expliquée (variation des $\hat{Y}_i = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$) sur la variation totale (variation des $Y_i = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$).
On exprime souvent r^2 en % **de variation expliquée par le modèle**.



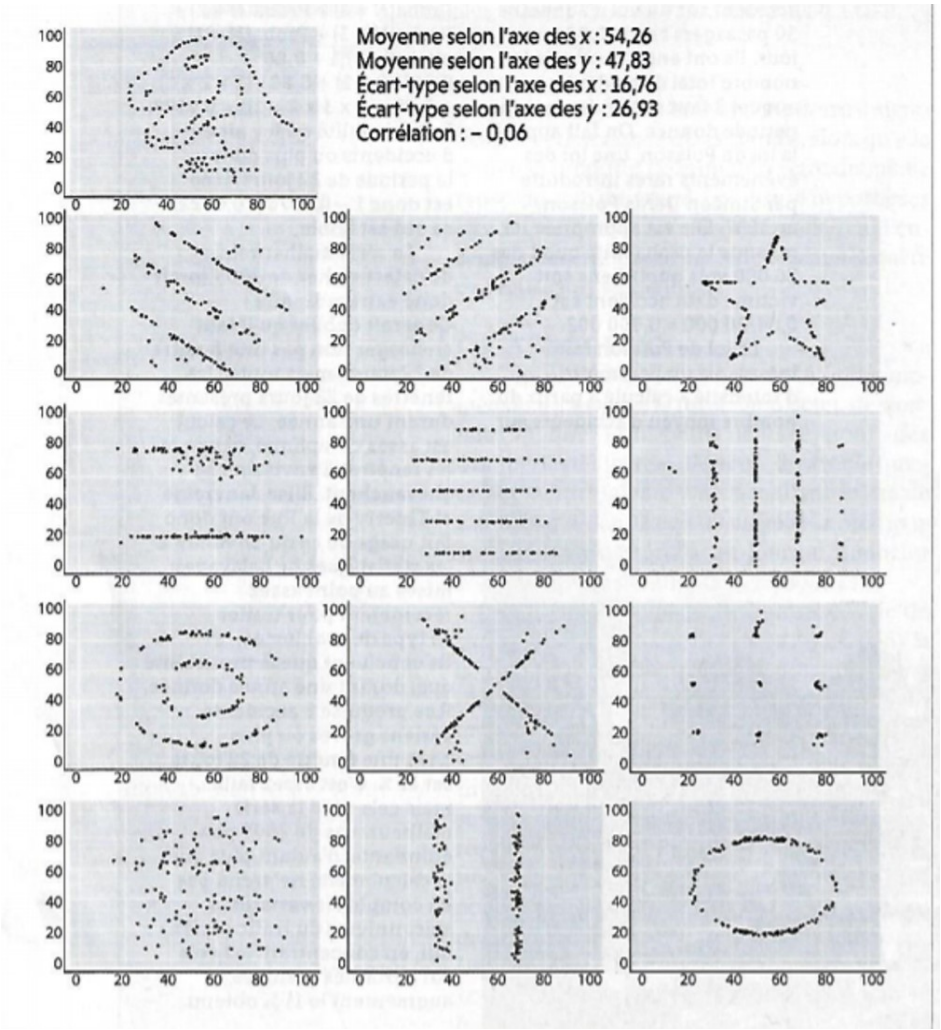
Suffit-il de regarder r^2 pour juger de la qualité d'un ajustement ?

NON ! Voici 4 exemples avec les mêmes valeurs de $r^2 = 62\%$
d'après R. Tomassone *et al.*, 1992, La régression, nouveaux regards sur une ancienne méthode statistique.



Il est capital de regarder le nuage de points

D'autres exemples jouets construits tous avec les mêmes paramètres statistiques.
Extrait d'un numéro du journal Pour la Science de Novembre 2017



Plan

Peut-on réaliser un test de corrélation linéaire dans le cadre de la régression linéaire ?

OUI,
celui-ci est fait automatiquement et correspond aussi au test d'égalité à 0 de la pente (affiché dans le résumé de la régression) appelé test de signification de la pente, qui répond à la question : "y a-t-il un effet significatif de X sur Y ?"

Peut-on tracer une droite de régression dans le cadre de la corrélation linéaire ?

NON.

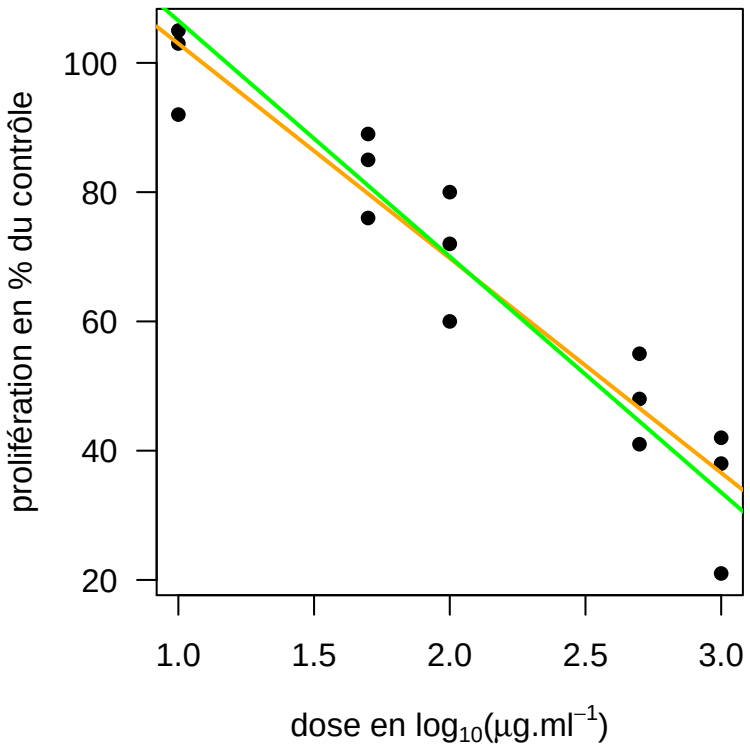
Le choix de la variable de contrôle (X) a un impact sur la droite de régression, donc si X et Y ont des rôles symétriques, aucune des 2 droites n'a de justification.

Erreur pourtant très courante !

On peut utiliser la régression linéaire
si X **est contrôlée** et Y **observée**,
ou dans un cadre élargi,
si Y **est une variable que l'on veut expliquer** (ou prédire) à partir de la **variable explicative** X.

Impact du choix de X et Y sur la droite de régression de Y en X

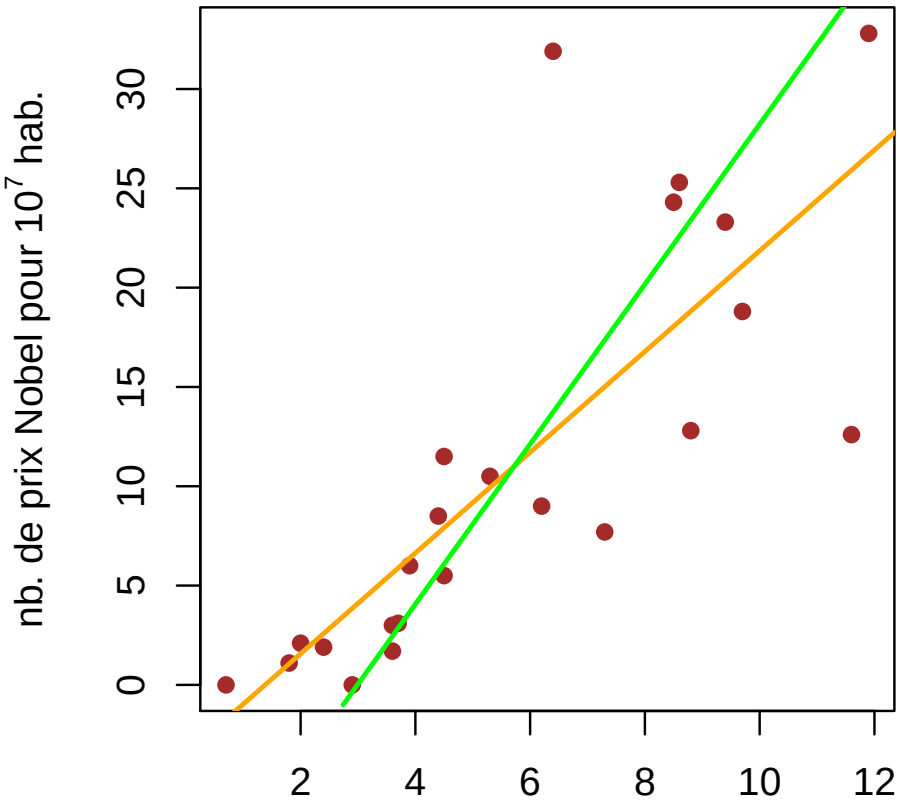
Comparaison des 2 droites sur notre exemple de base.



Soit Y la prolifération et
X la dose en log_{10}
Régression $Y = \alpha + \beta X + \epsilon$
Régression $X = \gamma + \delta Y + \epsilon$
Et plus la dispersion est grande et plus les droites diffèrent.

Il ne convient pas d'associer une droite de regression à un nuage de points dans la cadre de la corrélation linéaire

Reprenons l'exemple du cours sur la corrélation linéaire



Pourquoi choisirait-on plus l'une ou l'autre des deux droites de régression ?

Mieux vaut s'abstenir dans un tel cas !

Ici le graphe commet une double erreur car en sus les résidus ne respectent pas les conditions du modèle

Le modèle linéaire - régression multiple

Un modèle linéaire gaussien permet de modéliser l'**effet de plusieurs variables explicatives sur une variable à expliquer quantitative continue.**
Si les variables explicatives sont toutes quantitatives, on parle de régression multiple :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$
avec $\epsilon_i \sim N(0, \sigma)$

Partie déterministe : relation linéaire multiple
Partie stochastique : modèle gaussien
ϵ_i aléatoires, indépendants, suivant une loi normale de variance résiduelle σ^2 constante.

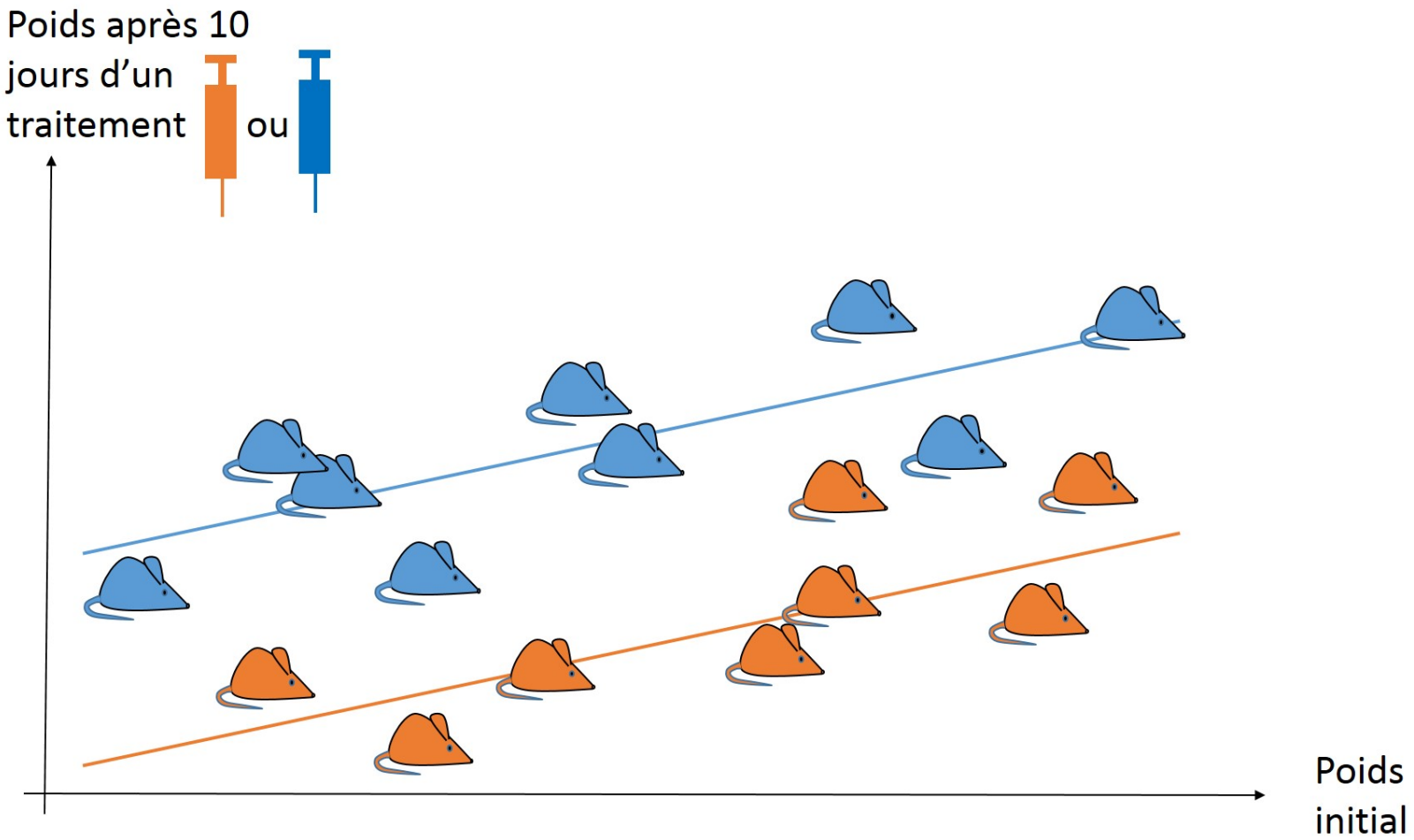
fonction `lm` dans **R**

Le modèle linéaire et ANOVA

Plus généralement un modèle linéaire gaussien permet de modéliser l'**effet de plusieurs variables explicatives quantitatives et/ou qualitatives sur une variable à expliquer.**
Il est alors nécessaire de coder les modalités des variables qualitatives par des variables muettes en utilisant $p-1$ variables muettes pour coder les p modalités d'un facteur (ou variable qualitative).
Exemple de codage d'un facteur traitement à 2 modalités :
traitement $= 0$ *si traitement A et traitement* $= 1$ *si traitement B*
Une analyse de variance peut être ainsi réalisée par ajustement d'un modèle linéaire gaussien.

fonction `lm` dans **R**

Le modèle linéaire - illustration schématique avec deux variables explicatives, une qualitative, une quantitative.



Le modèle non linéaire

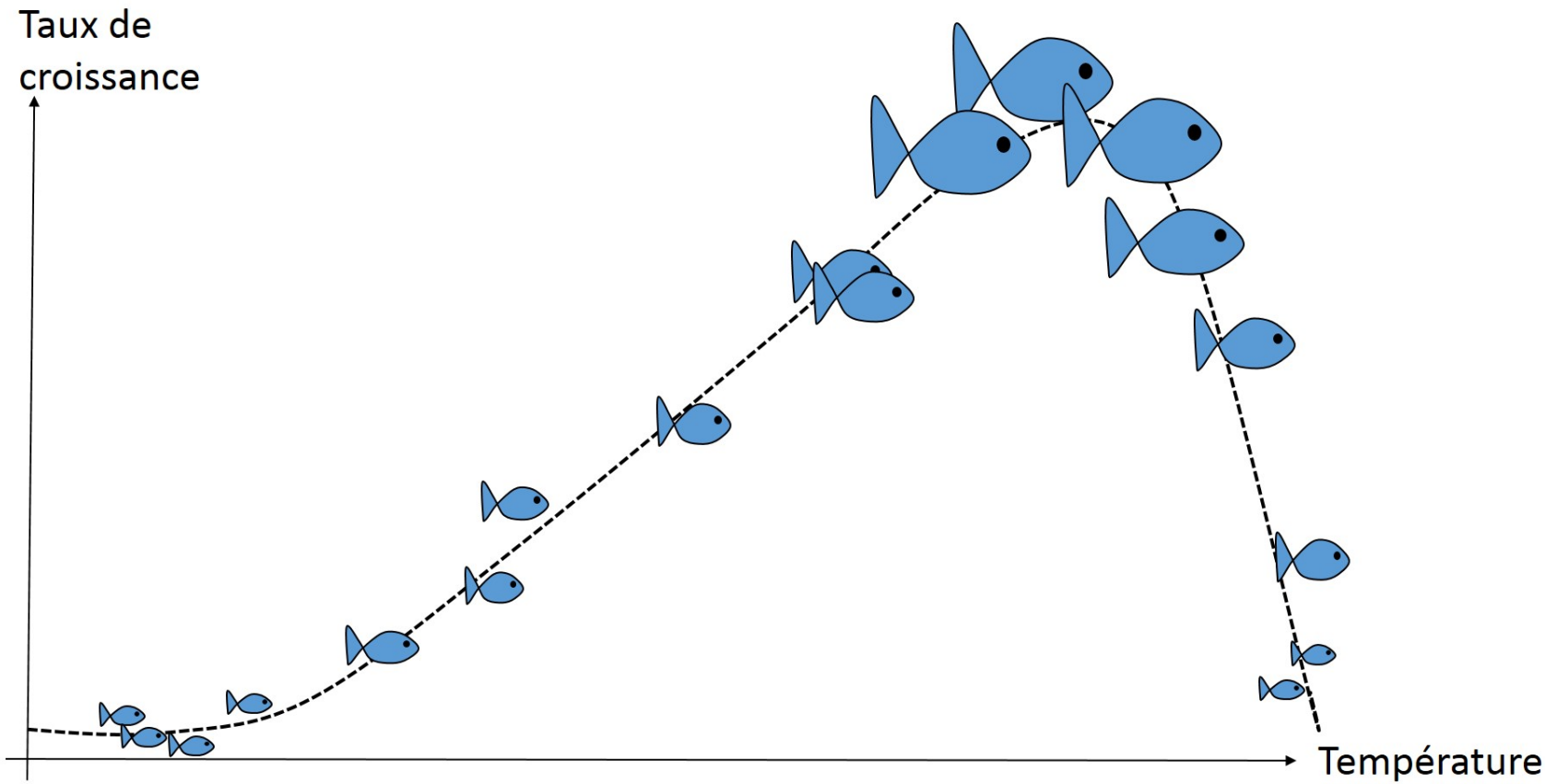
Un modèle est dit non linéaire si la variable à expliquer ne peut plus être exprimé comme une fonction linéaire des paramètres du modèle.

$$Y_i = f(X_i, \theta) + \epsilon_i$$
$$\text{avec } \epsilon_i \sim N(0, \sigma)$$

Ex. de modèle non linéaire : $Y_i = \alpha e^{\mu X_i} + \epsilon_i$ *avec* $\epsilon_i \sim N(0, \sigma)$
Ex. de modèle linéaire : $Y_i = \alpha + \beta X_i + \gamma X_i^2 + \epsilon_i$ *avec* $\epsilon_i \sim N(0, \sigma)$
Partie déterministe : fonction non linéaire des paramètres.
Partie stochastique : modèle gaussien.

fonction `nls` dans **R**

Le modèle non linéaire - illustration schématique

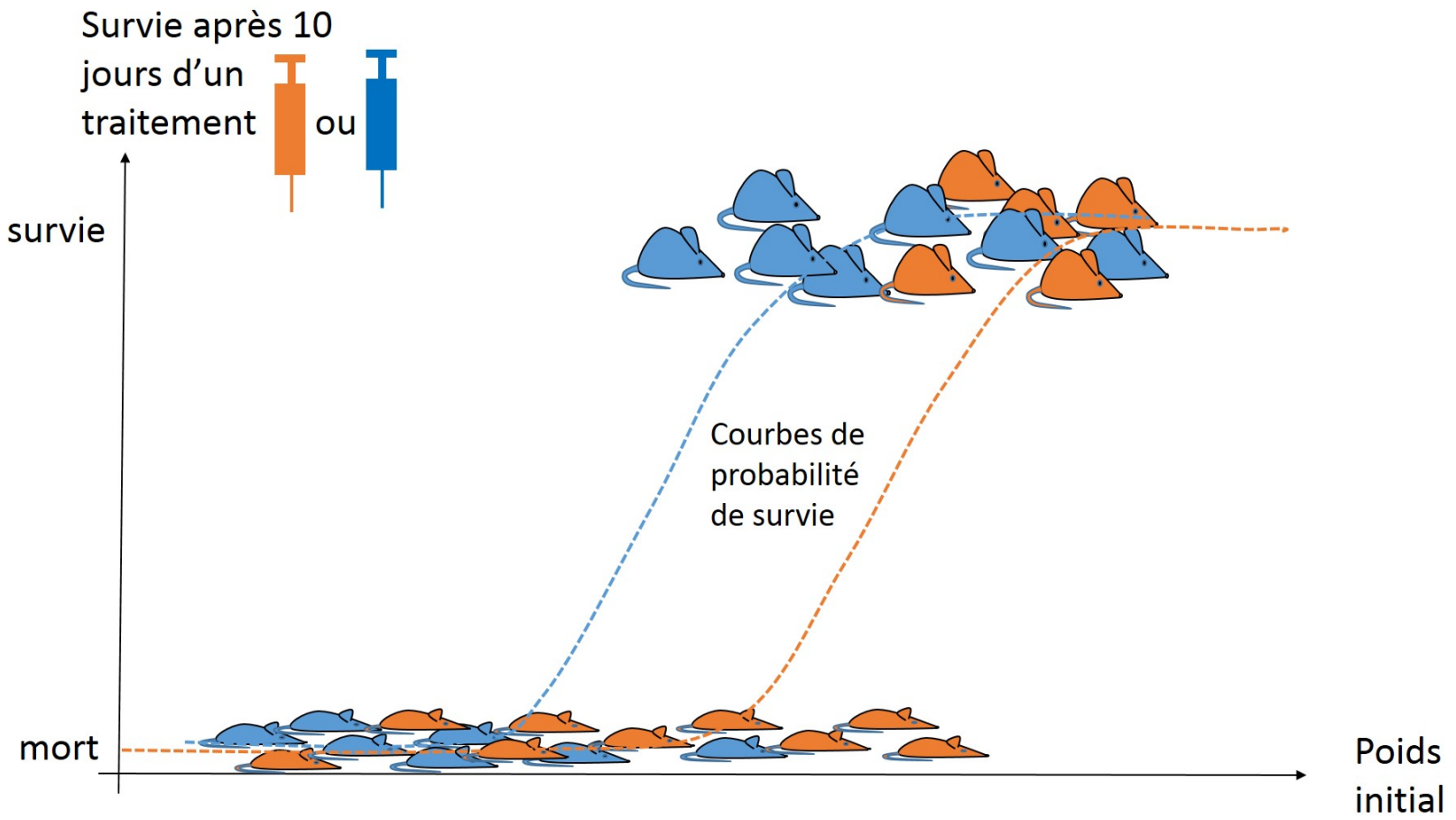


Le modèle linéaire généralisé

Un modèle linéaire généralisé permet de modéliser
l'**effet de plusieurs variables explicatives quantitatives et/ou qualitatives sur**
une variable à expliquer qualitative binaire
(ex. : malade / non malade)
ou une variable quantitative discrète
(ex. : nombre d'animaux par portée).
Partie déterministe : une transformation de la variable à expliquer (fonction de lien) est décrite par une fonction linéaire des variables explicatives .
Partie stochastique : le modèle n'est plus gaussien.

fonction `glm` dans **R**

Le modèle linéaire généralisé - illustration schématique

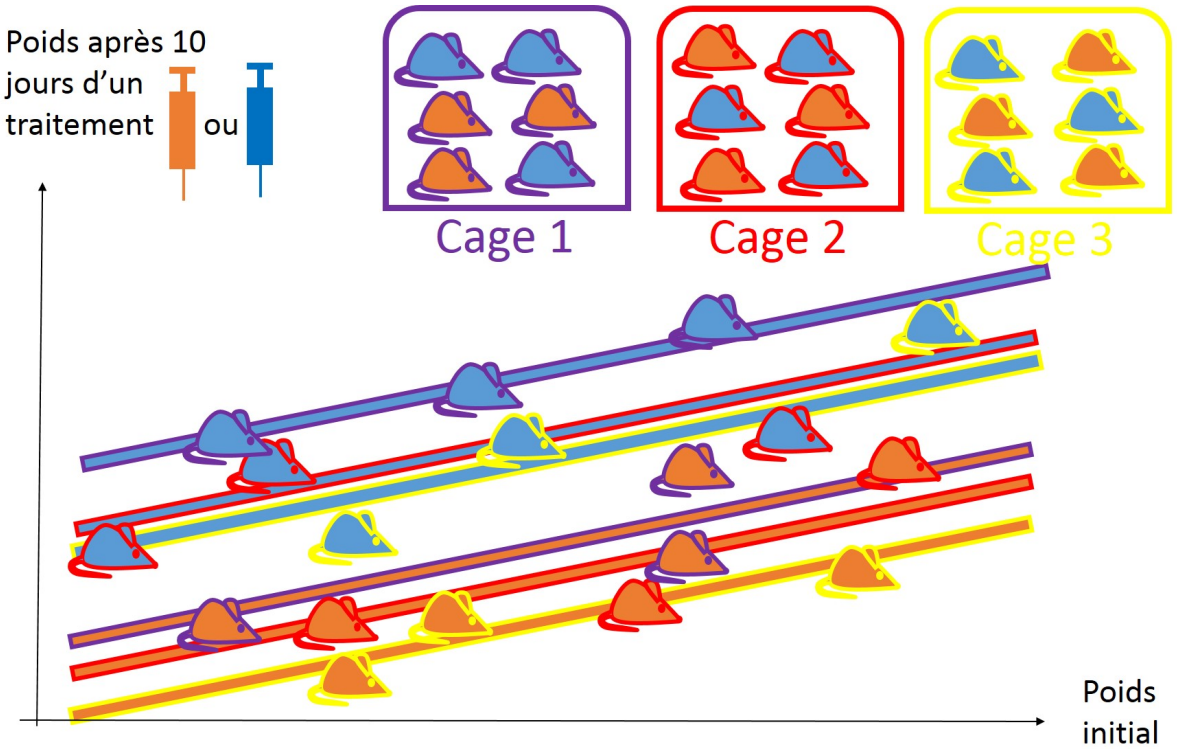


Le modèle linéaire mixte

Un modèle linéaire gaussien ne permet de prendre en compte que des facteurs (ou variables qualitatives) fixes, c'est-à-dire dont toutes les modalités d'intérêt sont observées. Lorsque seul un échantillon aléatoire des modalités d'un facteur sont observées, le **facteur est dit aléatoire** et l'on utilise alors un **modèle mixte** pour modéliser son effet sur la variable à expliquer.
Ex. : prise en compte d'un facteur "cage" ou "élevage"
Partie déterministe : linéaire .
Partie stochastique : modèle gaussien sur les ϵ_i et modèle gaussien sur les effets des facteurs aléatoires.

fonction `lmer` du package `lme4` dans **R**

Le modèle linéaire mixte - illustration schématique



Conclusion

Le modèle linéaire (avec ses extensions) est de loin l'outil le plus couramment utilisé au quotidien en statistique inférentielle.
Divers modules de formation concernant ce modèle linéaire et ses extensions sont proposés en formation complémentaire.
Programme détaillé sur
`http://www3.vetagro-sup.fr/ens/biostat/formcont.html`.