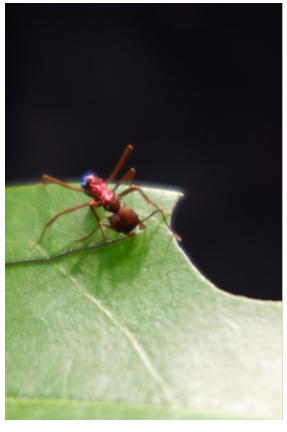
MoDoBio Modélisation des Données en Biologie

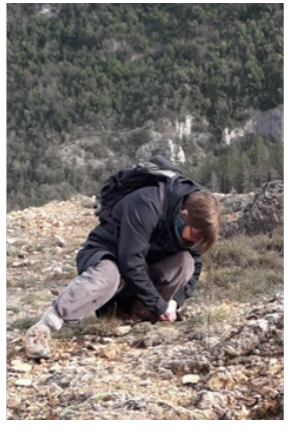
L3 Sciences de la Vie, parcours Biologie-Ecologie

Elodie Brunel MA et Céline Devaux BE

Université de Montpellier – Faculté des Sciences

2024-2025






Modèle Linéaire Simple

Objectifs

Données : deux variables mesurées / expérience

Etudier :

1. si les variables sont liées
2. quelle est la force du lien
3. si la variable d'intérêt peut-être prédite en observant uniquement l'autre

Exemple historique : les données de Galton

Le terme "regression" a été introduit par Francis Galton, chercheur britannique du 19e siècle, dans le célèbre article :
Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute* 15 : 246-63 (1886)

x : taille moyenne des deux parents
y : taille de l'enfant
Les données brutes sont disponibles :
http://www.randomservices.org/random/data/Galton.html

Expérience : on mesure la taille des enfants y et la taille des parents x au sein d'une même famille.

But : prédire la taille de l'enfant y en fonction de la taille moyenne des parents x.

$\longrightarrow$ pour un certain domaine de x, on s'attend à

$$y \approx \beta_0 + \beta_1 x.$$

Exemple : Abondance de truites et caractéristiques de l'habitat

On dispose de la mesure de la biomasse (en $g/100m^2$) de truites communes (Salmo trutta L.) dans 33 stations de la rivière Neste d'Aure dans le département des Hautes-Pyrénées et de diverses caractéristiques de l'habitat. (données disponibles dans Baran et al. Bull. Fr. Pêche Piscic. (1993) 331 : 321 -340.)

Objectifs : Caractériser la variabilité de la biomasse en fonction de différents prédicteurs environnementaux liés à l'habitat (Altitude, Température, Abris, Largeur, Profondeur de la rivière, etc) .

La Neste d'Aure à Saint Lary-Soulan (65)

Cadre

Notations

y : réponse ou variable d'intérêt ou variable à prédire
x : variable explicative ou prédicteur ou variable d'entrée
n : taille de l'échantillon

• x est contrôlée par la personne en charge de l'expérience, alors que y est une réponse observée.

• y peut dépendre de nombreux autres facteurs $\longrightarrow$ la relation linéaire ne peut pas relier exactement y à x : il y a une erreur.

x_i valeur observée de x sur la $i^{\text{ème}}$ expérience et
y_i valeur observée de y sur la $i^{\text{ème}}$ expérience.

Modèle Linéaire simple : un seul prédicteur

On veut identifier, estimer et valider un modèle de lien.

Différence avec la statistique descriptive : on modélise le terme d'erreur par des **variables aléatoires**.

Ici, on ne s'intéresse qu'aux dépendances **linéaires**.

On parle aussi de régression linéaire simple.

Exemple : la truite de la vallée d'Aure

Objectifs : On cherche à expliquer la variabilité de la biomasse (en $g/100\,m^2$) en fonction de la surface d'abris (en % de la surface de la station d'observation)

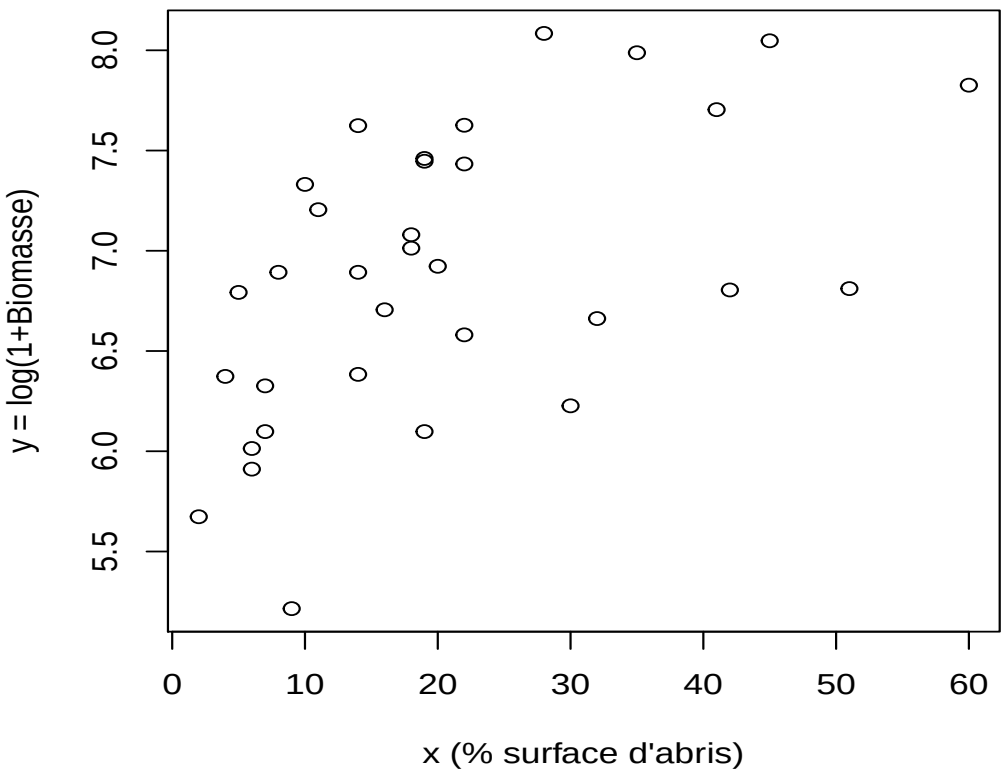
Expérience sur 33 stations de la Neste d'Aure

station	x_i : % abris	y_i : biomasse (en $g/100\,m^2$)
N1	16	816
N2	7	444
N3	19	444
N4	22	1690
N5	5	890
...	...	...

La biomasse augmente t'elle avec la surface d'abris ? Si oui, de combien ? peut-on le quantifier ?

Première étape : nuage de points et coefficient de corrélation

Le nuage de points est la représentation graphique des points de coordonnées cartésiennes (x_i, y_i) , $i = 1, \ldots, n$. Le nombre de points est noté n, c'est le nombre d'individus observés, ici 1 individu $=$ 1 station et $n = 33$ stations. Le coefficient de corrélation r (ou ρ)$= 0,55$.



Deuxième étape : Modélisation statistique

On fait une hypothèse sur la forme de la dépendance : c'est un **modèle**.

Y est une variable aléatoire qui dépend de x par

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad \text{où}$$

- $\epsilon_i =$ erreurs, variables aléatoires non observées, supposées indépendantes, identiquement distribuées de loi normale $\mathcal{N}(0; \sigma_\epsilon^2)$
- les paramètres β_0 et β_1 sont inconnus et la variance σ_ϵ^2 est inconnue.

Remarques importantes

Dans ce modèle, on a pour $i = 1, \cdots, n$,

1. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(\epsilon_i) = \beta_0 + \beta_1 x_i$;
2. $\mathrm{Var}(Y_i) = \mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$.

Attention, dans ce modèle, la variance σ_ϵ^2 est identique à chaque expérience. On parle d' **homoscédasticité**. Il existe des modèles hétéroscédastiques (pas vus dans ce cours).

Définition

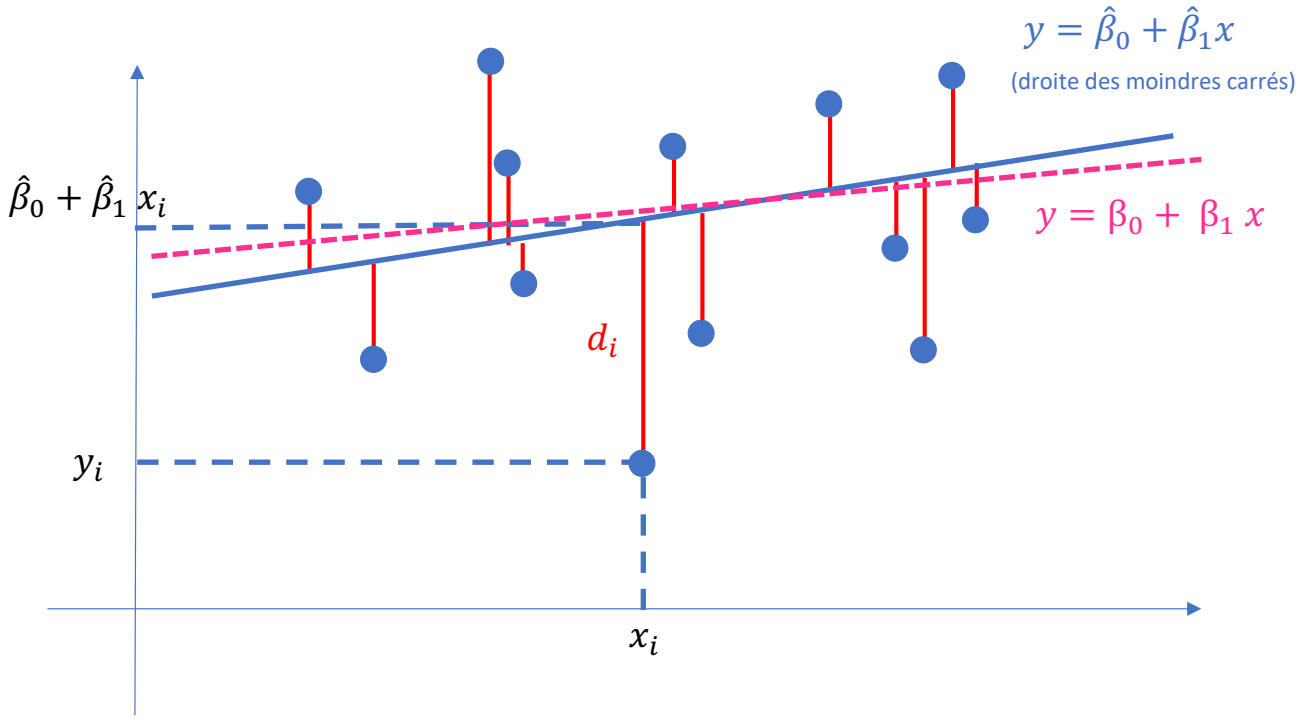
$\beta_0 + \beta_1 x$ s'appelle la **réponse moyenne** sachant x, ou espérance de la réponse Y sachant x : C'est la réponse moyenne de Y que l'on s'attend à observer, étant donnée une valeur de x.

Une fois le modèle posé, notre but est :

Trouver la droite qui prédit y au mieux connaissant x.

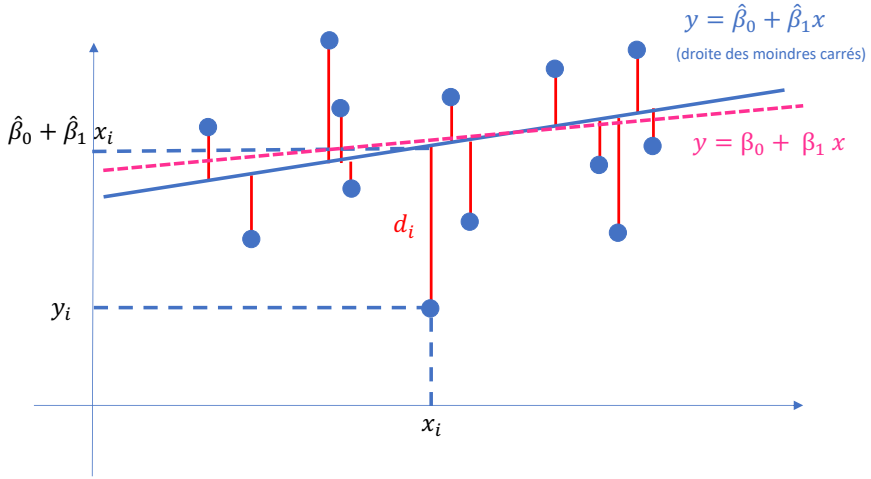
$\longrightarrow$ i.e. estimer β_0 et β_1.

Pour cela, il faut se donner un critère : c'est le critère des **moindres carrés**.



Principe de la méthode des moindres carrés

La droite des moindres carrés est celle qui minimise la somme des carrés des écarts entre les points du nuage et cette droite (traits verticaux rouge sur le schéma).



$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

est minimum.

Estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$

Quelques notations :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{et} \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Ici, les valeurs des paramètres $\hat{\beta}_0$ et $\hat{\beta}_1$ pour lesquelles la somme des carrés $\sum_{i=1}^{n} d_i^2$ est minimale sont données par :

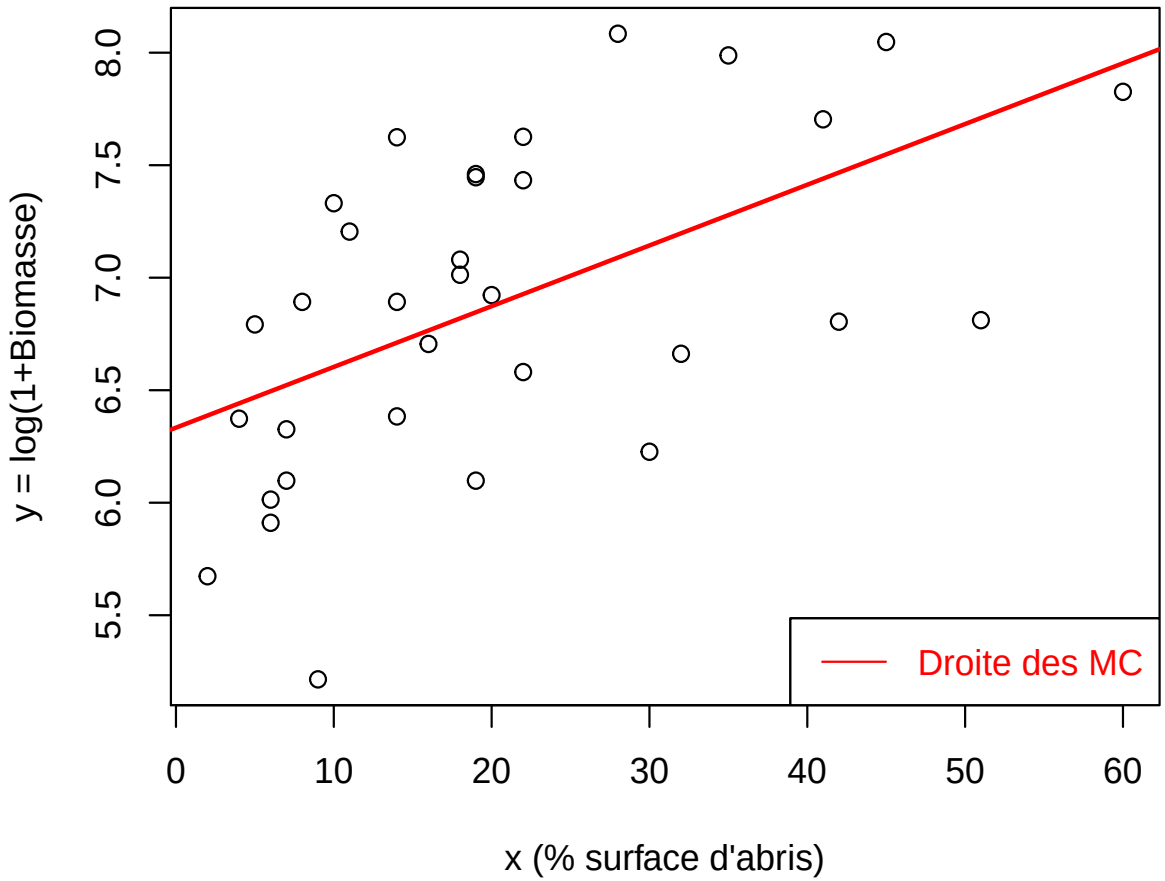
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

L'équation de la droite des moindres carrés (ou droite de régression de y en x) est $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

$\longrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, est appelée **valeur ajustée** par le modèle.

$\longrightarrow \hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ est la **meilleure prédiction** de la réponse Y sachant une nouvelle valeur x_{new}.

Exemple sur les données de la truite de la vallée d'Aure



$\sum_{i=1}^{n} d_i^2 = 11,692 \quad \hat{\beta}_0 = 6,33 \quad \hat{\beta}_1 = 0,027$

Equation de la Droite des Moindres Carrés : $y = 6,33 + 0,027x$.

Exemple sur les données de la truite de la vallée d'Aure

Détails des calculs :
x : % surface des abris
$y = \log(1 + \textit{biomasse})$ (nous reviendrons sur le choix de ce changement de variable)

Grâce aux 4 quantités clés (calculées à partir des données) :

$$\bar{x} = 20,485 \quad \bar{y} = 6,886 \quad S_{xy} = 182,576 \quad S_{xx} = 6764,242$$

on peut déterminer les valeurs de $\hat{\beta}_1$, puis $\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{182,576}{6764,242} \simeq 0,02699 \quad \hat{\beta}_0 = 6,886 - 0,02699 \times 20,485 \simeq 6,33$$

D'où l'équation de la Droite des MC : $y = 6,33 + 0,027x$.

Deux sources de variabilité de Y dans le modèle

1. La variabilité due à $x \longrightarrow$ celle due au "modèle"

2. La variabilité due aux erreurs ϵ non-observées, à x fixé.
$\longrightarrow$ On cherche à estimer la variance σ_ϵ^2 des ϵ_i qui ne sont pas observés :

Résidus

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{⚠} \quad e_i \neq \epsilon_i$$

Somme des Carrés des Résidus

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

Remarque : SSR est aussi égale à $\sum_{i=1}^{n} d_i^2$ (la somme des carrés des écarts des points à la droite des MC)

Estimateur s^2 de la variance σ_ϵ^2

Dans le modèle, $\sigma_\epsilon^2 = \mathrm{Var}(\epsilon_i)$.

$\underbrace{\epsilon_i = Y_i - \beta_0 - \beta_1 x_i}_{\text{non observés}} \longrightarrow e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

$\longrightarrow$ on estime la variance des résidus e_i.

On remarque que $\sum_{i=1}^n e_i = 0$ donc la moyenne des résidus $\bar{e} = 0$.

Estimation de la variance σ_ϵ^2 par

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SSR}{n-2}$$

Exemple des truites. $SSR \approx 11,692$ et $n = 33$ donc

$$s^2 = \frac{11,692}{31} \simeq 0,377$$

i.e. l'estimation de l'écart-type estimé est $s = \sqrt{0,377} \simeq 0,614$.

Un point important

Le modèle est $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, où β_0, β_1 sont inconnus. La droite des moindres carrés $y = \hat{\beta}_0 + \hat{\beta}_1 x$ fournit une estimation de la réponse moyenne, qui est inconnue, pour des valeurs de x dans le même intervalle que les données.

Dans l'exemple, les données x (% surface d'abris) s'étalent entre 2 et 60 $\longrightarrow$ pour $x^* = 10\%$, on peut donner la prédiction de y : $\hat{y} = 6,33 + 0,027 \times 10 = 6,60$.

Trois Questions.
1. Quelle est la force du lien entre la réponse Y et la variable explicative x ?
2. β_1 est-il proche de $\hat{\beta}_1$?
3. Quelle est la précision de la prédiction $\hat{y}_{new}$ pour une nouvelle valeur x_{new} ?

Force du lien linéaire : la décomposition de la variabilité

Réponse	$=$	Valeur expliquée par x	$+$	résidu
y_i	$=$	$\hat{y}_i$	$+$	e_i.

La quantité $S_{yy} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ mesure la variabilité de y autour de $\bar{y}$: on l'appelle aussi la **somme des carrés totale** et on la notera SST, elle se décompose :

Théorème (de décomposition de la "Variance")

Variabilité de y (Totale)	$=$	Variabilité expliquée par le Modèle	$+$	Variabilité Résiduelle
SST	$=$	SSM	$+$	SSR

avec $SST = S_{yy}$; $SSM = \frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

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

Force du lien : le coefficient de détermination r^2

$$SST = SSM + SSR \text{ qui s'écrit aussi : } S_{yy} = \frac{S_{xy}^2}{S_{xx}} + SSR.$$

En divisant de chaque côté de l'égalité par S_{yy}, on fait apparaître les proportions de variabilité expliquée par le Modèle et Résiduelle :

$$1 = \underbrace{\frac{S_{xy}^2}{S_{xx}S_{yy}}}_{\text{"expliquée" par le modèle}} + \underbrace{\frac{SSR}{SST}}_{\text{résiduelle}}.$$

La **proportion de la variabilité** de Y **"expliquée"** par le modèle est notée :

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSM}{SST}.$$

en effet on reconnaît le carré du coefficient de corrélation linéaire de Pearson r ou ρ. Le coefficient r^2 est aussi appelé coefficient de détermination.

Force du lien : retour à l'exemple de la truite de la vallée d'Aure

A partir des quantités SSM et SSR, on peut calculer le coefficient de détermination et évaluer la force du lien linéaire entre la variable d'intérêt $y = \log(1 + \textit{biomasse})$ et la variable explicative ou prédicteur x % surface d'abris :

$$r^2 = \frac{SSM}{SST} = \frac{SSM}{SSM + SSR} = \frac{4,928}{4,928 + 11,692} \simeq 0,2965 \text{ soit } \approx 30\%$$

Conclusion : Le % de surface d'abris explique seulement 30 % de la variabilité de la biomasse. Ceci n'est pas très étonnant car on se doute que d'autres facteurs entrent en jeu.

Construction d'intervalles de confiance et tests d'hypothèse

Précision de l'estimation fournie par les estimateurs des moindres carrés : on peut construire des intervalles de confiance des paramètres β_0 et β_1. Pour cela, on a :

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad \text{et} \quad \mathbb{E}(\hat{\beta}_1) = \beta_1$$

Puis, on estime leur variance :
Les variances des estimateurs des moindres carrés sont

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{xx}}, \quad \mathrm{Var}(\hat{\beta}_0) = \sigma_\epsilon^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\Big)$$

Mais σ_ϵ^2 est inconnu et doit être estimé par $s^2 = \frac{SSR}{n-2}$. D'où,

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}}, \quad \widehat{\mathrm{Var}}(\hat{\beta}_0) = s^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\Big)$$

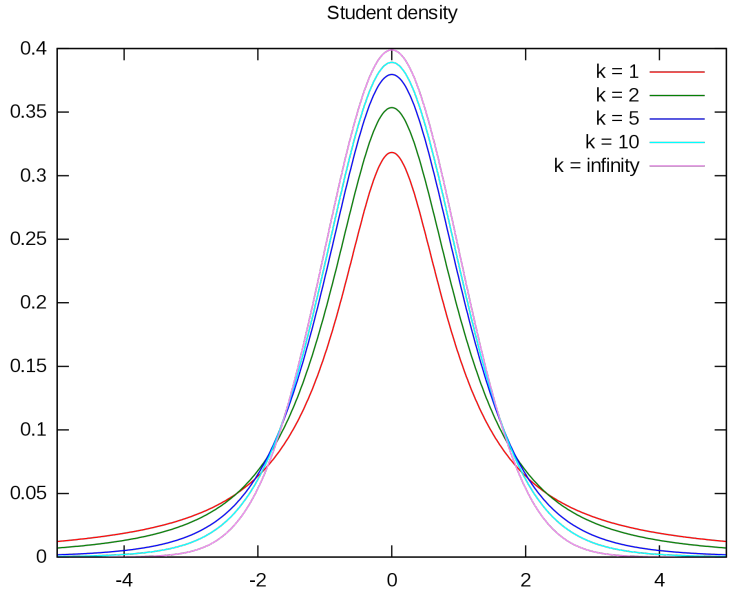
Loi de probabilité (ou distribution) des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_0$ centrés et réduits

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}}$$

suit une loi de Student à $n-2$ d.d.l.

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

suit une loi de Student à $n-2$ d.d.l.



(d'après Wikipédia)

Problèmes d'inférence importants

- Tester des hypothèses,
- Construire des intervalles de confiance
- Faire des prédictions

dans le contexte de la régression linéaire simple.

Inférence sur la pente β_1

Tester la nullité de la pente β_1

- On souhaite tester $\mathcal{H}_0 : \beta_1 = 0$ contre $\mathcal{H}_1 : \beta_1 \neq 0$ (ce qui se traduit par : la variable x a t'elle un lien (linéaire) significatif avec la variable d'intérêt y ?)
- La statistique de test sous $\mathcal{H}_0$ est :

$$T_1 = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$$

  Elle suit une loi de Student à $n-2$ d.d.l sous $\mathcal{H}_0$.
- Règle de décision : on rejette $\mathcal{H}_0$ pour un risque de première espèce α si la valeur de la statistique $T_{1,\text{obs}}$ qui a été observée sur les données n'est pas dans l'intervalle $[-t_{1-\alpha/2} \; ; \; +t_{1-\alpha/2}]$.

  où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ d.d.l.

Remarque Ce test permet de tester l'existence d'un lien linéaire entre x et y.

Exemple sur les données de la truite de la vallée d'Aure

- Hypothèse nulle $\mathcal{H}_0 : \beta_1 = 0 \Longleftrightarrow$ pas de lien (linéaire) entre la biomasse et le % surface d'abris.
- On a vu $\hat{\beta}_1 = 0,02699$ et $s = 0,614$ et $S_{xx} = 6764,242$.

  Statistique de test observée : $T_{1,\text{obs}} = \frac{0,02699}{0,614/\sqrt{6764,242}} = 3,615$

  Avec $\alpha = 5\%$, $n - 2 = 31$, $t_{0.975} = 2.039$

- Règle de décision : $\longrightarrow$ $T_{1,\text{obs}} = 3,615 > 2.039$ DONC on rejette $\mathcal{H}_0$.

Conclusion : il y a bien un lien significatif entre le % surface d'abris et la biomasse (même s'il ne suffit pas à lui seul à expliquer tout la variabilité de la biomasse !).

Intervalle de confiance sur la pente

Au niveau $(1-\alpha)$, l'I.C. de β_1 est

$$\left[\hat{\beta}_1 - t_{1-\alpha/2}\frac{s}{\sqrt{S_{xx}}} \; ; \hat{\beta}_1 + t_{1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right]$$

Exemple des données. L'intervalle de confiance pour β_1 de niveau 95% est $[0,0118; 0,0422]$

Remarque : S'il ne contient pas la valeur 0 alors on rejette $\mathcal{H}_0 : \beta_1 = 0$ au profit de $\mathcal{H}_1 : \beta_1 \neq 0$.

Test de Fisher et analyse de la variance

Table d'analyse de la variance et test de Fisher global

```
Analysis of Variance Table

Response:   logBiomTot
            Df   Sum Sq   Mean Sq   F value   Pr(>F)
Abris        1    4.928    4.9280    13.066   0.001052 **
Residuals   31   11.692    0.3772
---
```

Source	Degrés de liberté	Somme des Carrés	Carrés Moyens	Statistique de test	Probabilité critique (p-value)
Modèle	p	$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\frac{SSM}{p}$	$F_{obs} = \frac{SSM/p}{SSR/(n-p-1)}$	$P(F_{p,n-p-1} > F_{obs})$
Résidus	$n-p-1$	$SSR = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$\frac{SSR}{n-p-1}$		
Total	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$\frac{SST}{n-1}$		

avec p le nombre de variable(s) explicative(s) dans le modèle : $p = 1$ avec une variable explicative

Test de Fisher "global"

Le test de Fisher teste la significativité "globale" du modèle (cf. modèle multivarié) :

$\mathcal{H}_0 : Y = \beta_0 + \varepsilon$ contre $\mathcal{H}_1 : Y = \beta_0 + \beta_1 x + \varepsilon$
$\iff$ (dans le cas d'une seule variable explicative x) :
$\mathcal{H}_0 : \beta_1 = 0$ contre $\mathcal{H}_1 : \beta_1 \neq 0$
Sous $\mathcal{H}_0$, on a :

$$F = \frac{SSM/p}{SSR/(n-p-1)} \sim F_{p,n-p-1}$$

Loi de Fisher à p et $n-p-1$ degrés de liberté (d.d.l) où p est le nombre de variable(s) explicative(s)

Remarque : Pour une seule variable explicative on peut montrer que $F = T_1^2$ le test de Fisher et le test de Student de nullité de la pente sont équivalents.


cela ne sera plus vrai dans le cas multivarié.

Prédiction et Intervalle de prédiction

Prédiction et Intervalle de prédiction

Pour une nouvelle valeur de $x = x_{new}$ donnée, on prédit la valeur de la réponse Y par

$$\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

(à ne pas confondre avec une valeur ajustée pour une valeur x_i observée (de l'échantillon) : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$)

L'intervalle de prédiction tient compte de l'imprécision que l'on a sur l'estimation des paramètres β_0, β_1 et σ^2 :

Intervalle de prédiction au niveau $1 - \alpha$ pour une nouvelle valeur x_{new}

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} \pm t_{1-\alpha/2} \times s \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

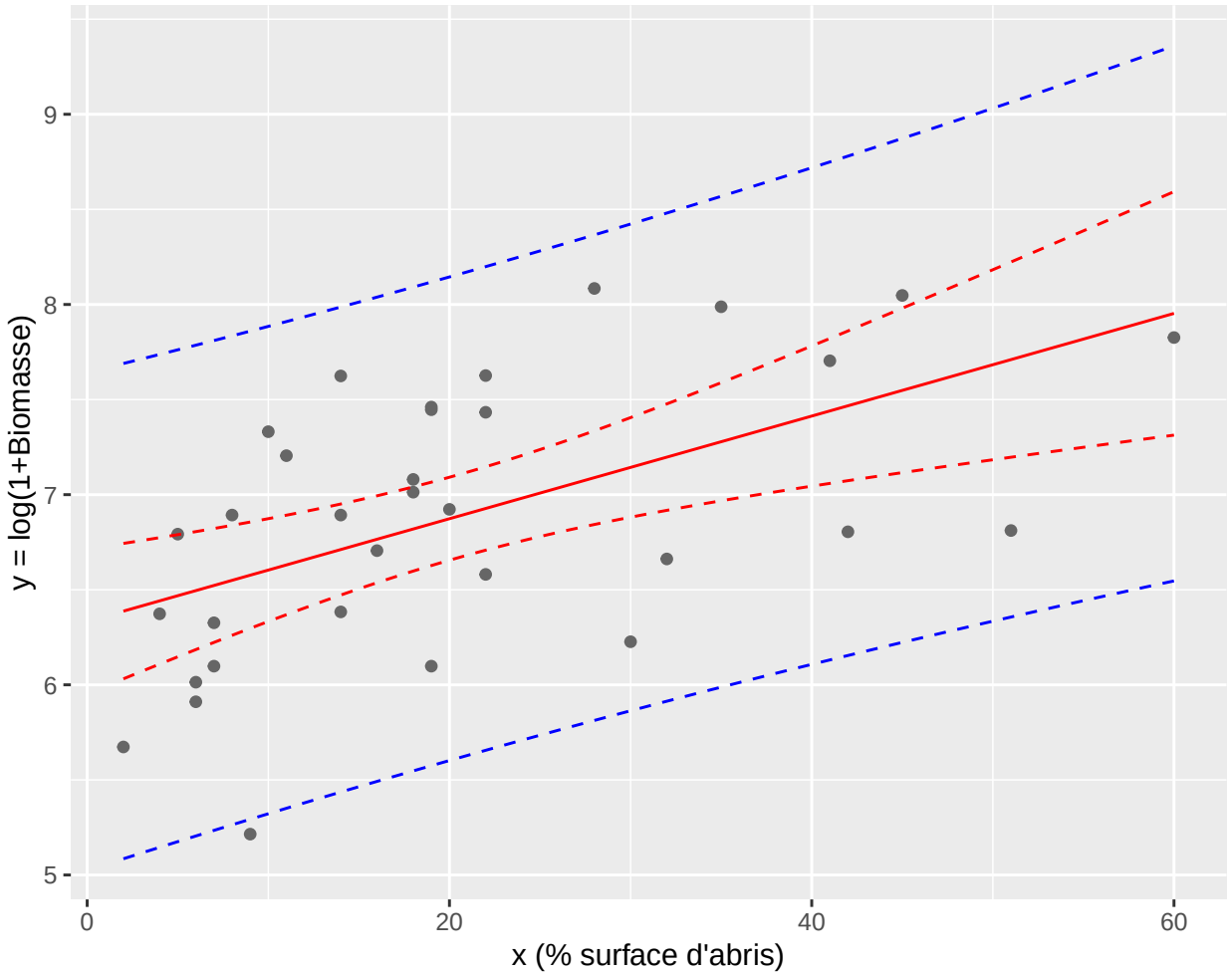
à ne pas confondre avec [1] :

Intervalle de Confiance de la réponse moyenne étant donné $x = x_i$

$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{1-\alpha/2} \times s \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$$

1. Les formules ne sont pas à connaître par cœur.

Intervalle de prédiction *v.s* Intervalle de confiance de la réponse moyenne



l'intervalle de prédiction est TOUJOURS plus large que l'Intervalle de Confiance de la réponse moyenne

Quelques remarques sur le modèle de régression linéaire simple

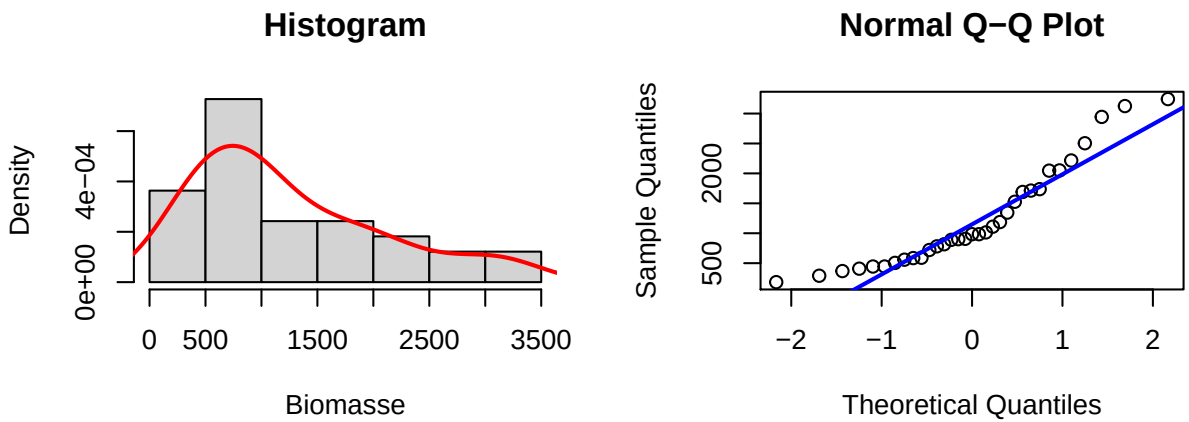
Quelques remarques sur le modèle de régression linéaire simple

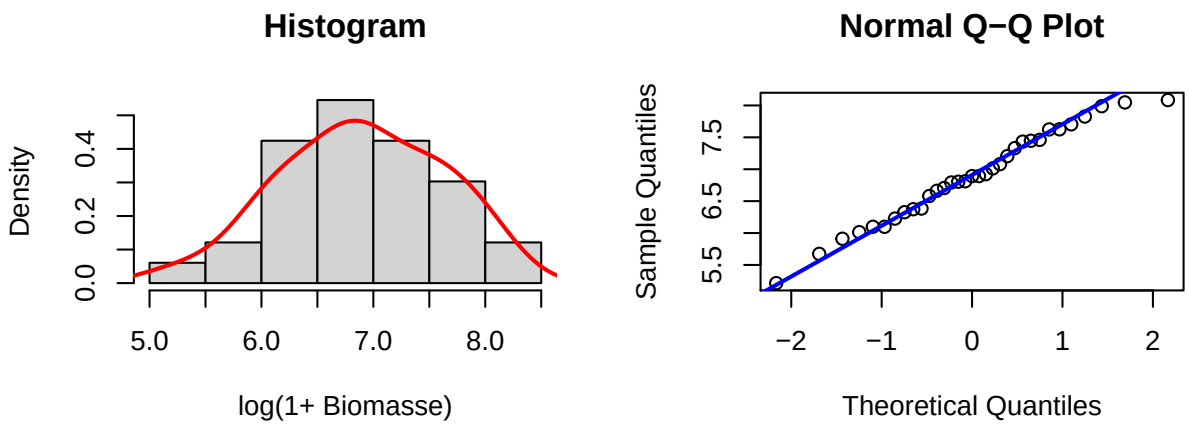
Toutes les analyses que nous venons de faire sont valides à condition que les hypothèses suivantes soient vérifiées :

- la relation entre la réponse et la variable explicative est linéaire
- les erreurs ϵ_i sont indépendantes
- la variance de l'erreur ne dépend pas de l'expérience (elle est constante)
- l'erreur ϵ_i suit une loi normale.

Transformation préliminaire de la variable réponse Y

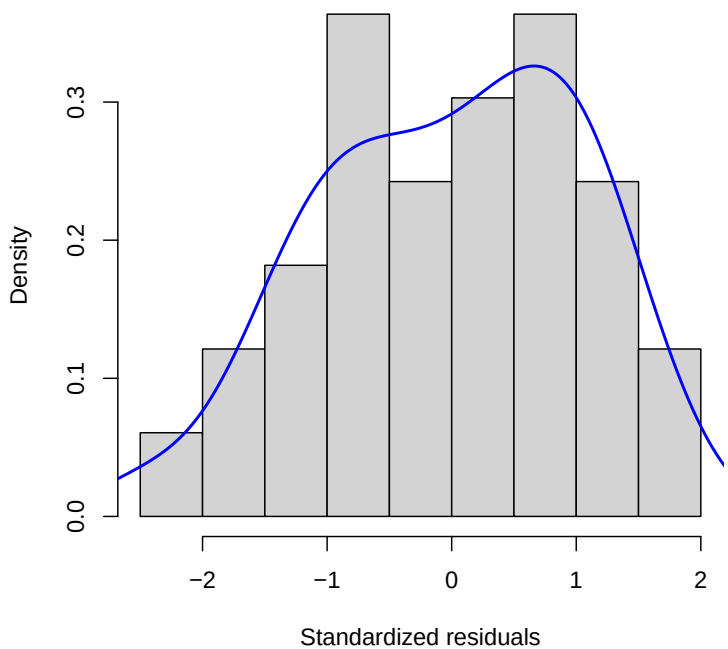
Avant de faire l'analyse : On trace l'histogramme et le QQ-plot pour "valider" ou pas la normalité de la variable y : Biomasse. On peut faire un changement de variable pour se rapprocher de la "normalité" .

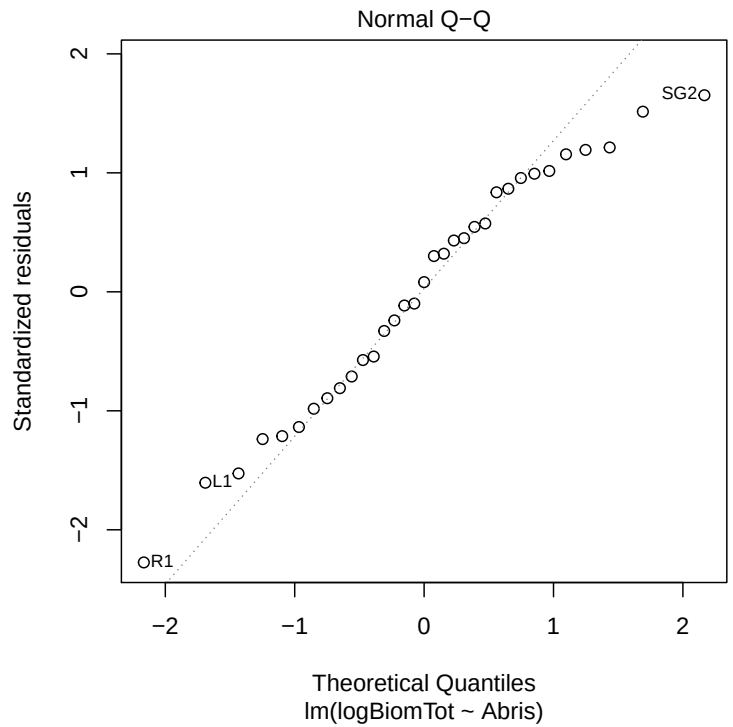




Vérification *a posteriori* de la normalité des résidus

Après l'estimation, on vérifie la normalité des résidus $e_i = y_i - \hat{y}_i$ ou des résidus "standardisés" ou "studentisés" (c'est-à-dire qu'on les a réduits en divisant par leur écart-type) :





On peut compléter par un test de Shapiro (cf. TP)

Recherche de l'absence/présence de tendance dans les résidus

Après l'estimation, si les résidus $e_i = y_i - \hat{y}_i$ présentent une "tendance" c'est qu'ils ne peuvent à eux seuls expliquer toute la variabilité de la variable Y. On peut représenter les résidus e_i (ou résidus studentisés) en fontion des valeurs prédites $\hat{y}_i$ pour le détecter