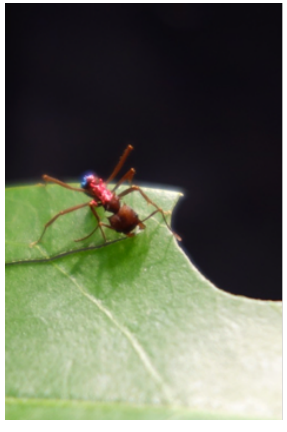
MoDoBio Modélisation des Données en Biologie

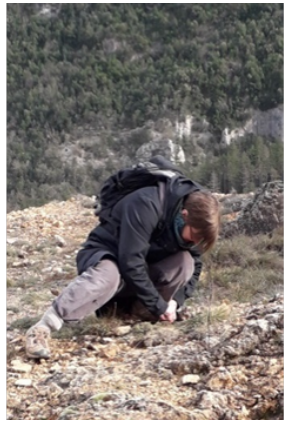
L3 Sciences de la Vie, parcours Biologie-Ecologie

Elodie Brunel MA et Céline Devaux BE

Université de Montpellier – Faculté des Sciences

2024-2025






Modèle Linéaire Multiple

(Extension au cas de plusieurs variables explicatives)

Retour en vallée d'Aure ...

Nous avons mis en évidence avec la modélisation linéaire simple que :

- il existe un lien significatif entre la biomasse et la variable explicative % surface d'abris
- le % surface d'abris explique 30% de la variabilité de la biomasse. Elle ne suffit pas à elle seule à expliquer toute la variabilité de la biomasse.

Nous voulons enrichir le modèle linéaire en introduisant d'autres variables explicatives disponibles :
Altitude, Température, Conductivité électrique, Pente, Largeur, Profondeur, Vitesse de Fond, Vitesse de Surface, Abris, Densité d'Invertébrés, Module, DebitE

Description des $p = 12$ variables explicatives disponibles

- Altitude (en m)
- Température moyenne du mois le plus chaud (en ^oC).
- Conductivité électrique (en $\mu S/cm$)
- Pente (en %) de la ligne d'eau,
- Largeur moyenne (en m) mesurée à partir de transects positionnés tous les 6 m,
- Profondeur moyenne (en m),
- Vitesse de Fond (en m/s) mesurée à l'aide d'un courantomètre,
- Vitesse de Surface (en m/s) calculée en chronométrant le temps d'écoulement d'un colorant,
- Abris % de surface d'abris en effectuant le rapport de la surface occupée par les abris sur la surface totale de la station.
- Densité d'Invertébrés (individus $/0,1\,m^2$),
- Module (en m^3/s) représente la quantité totale d'eau circulant pendant une année moyenne sur un tronçon de rivière.
- DebitE (en % du Module) est le débit du cours d'eau à l'étiage (période où le débit est le plus faible)

Modèle Linéaire Multiple (ou Multivarié)

Soit Y une variable aléatoire qui dépend de p variables explicatives $x_1, \cdots, x_p$ par :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \, , i = 1, \cdots, n.$$

où

- Y_i est la variable réponse de l'individu i
- x_{ij} est la valeur observée de la variable x_j pour l'individu i
- les erreurs ε_i sont inobservées, mais supposées indépendantes et de loi $\mathcal{N}(0, \sigma_\epsilon^2)$.
  $\beta_0, \beta_1, \cdots, \beta_p$ sont les coefficients inconnus du modèle.

Le modèle linéaire s'écrit de façon matricielle :

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}:n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{\mathbf{X}:n\times p} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta:p\times 1} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon:n\times 1}$$

Les estimateurs des moindres carrés minimisent la somme des carrés : $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$

et ils s'obtiennent *via* des calculs matriciels (omis ici ...). Les logiciels de statistique fournissent les valeurs numériques des estimateurs :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

On retrouve (comme dans le modèle linéaire simple), la formule de décomposition de la variabilité de la réponse :

$$SST = SSM + SSR$$

avec :
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$; $SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$; $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ et $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

Test de Fisher "global"

Une régression multiple s'accompagne toujours d'une analyse de variance (ANalyse Of VAriance = ANOVA). On cherche si l'ensemble des variables explicatives influent de façon significative sur la variable réponse.

$\mathcal{H}_0$: "Aucune des variables n'influe significativement sur la réponse" contre $\mathcal{H}_1$: "Au moins une variable a un lien significatif"
ce qui se traduit par :
$\mathcal{H}_0 : \beta_1 = \cdots = \beta_p = 0$ contre $\mathcal{H}_1$: au moins un $\beta_j \neq 0$
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 nombre de variables explicatives

Remarque : Ce test ne permet pas de décider quels sont les coefficients dans le modèle de régression qui contribuent de façon significative à la réponse : test "global".

Test de Fisher "global"

Régle de décision : On rejette $\mathcal{H}_0$ pour un risque de première espèce α si la valeur observée F_{obs} de la F-statistique sur les données est telle que :

$$F_{obs} > \text{quantile d'ordre } \alpha \text{ d'une loi } F_{p,n-p-1}$$

ou de façon équivalente si la probabilité critique ou la p-value associée à F_{obs} est inférieure à α.

Exemple : Test de Fisher pour le modèle de la biomasse avec 12 coefficients de l'habitat

$$Y = \beta_0 + \beta_1 \texttt{Altitude} + \beta_2 \texttt{Temp} + \beta_3 \texttt{Cond} + \beta_4 \texttt{Pente} + \beta_5 \texttt{Larg} + \beta_6 \texttt{Prof} + \beta_7 \texttt{V.Fond} + \beta_8 \texttt{V.Surf} + \beta_9 \texttt{Abris} + \beta_{10} \texttt{DensInv} + \beta_{11} \texttt{Module} + \beta_{12} \texttt{DebitE} + \epsilon$$

```
Residual standard error: 0.4193 on 20 degrees of freedom
Multiple R-squared:  0.7884,Adjusted R-squared:  0.6614
F-statistic: 6.209 on 12 and 20 DF,  p-value: 0.0001867
```

Table ANOVA avec $p = 12$ nombre de variables explicatives

Source	Degrés de liberté	Somme des Carrés	Carrés Moyens	F - statistique de test	p-value
Modèle	12	$SSM = 13,103$	$\frac{SSM}{12}$	$F_{obs} = \frac{13,103/12}{3,517/20} = 6,209$	0,00019
Résidus	20	$SSR = 3,517$	$\frac{SSR}{20}$		

On rejette l'hypothèse $\mathcal{H}_0$: "Aucune des variables n'influe sur la réponse" .

Remarque :

$r^2 = 0,7884$ et $r^2_{ajusté} = 0,6614$ "corrigé" lorsque $p > 1$.

Exemple : Estimation et test de Student des 12 coefficients de l'habitat

$$Y = \beta_0 + \beta_1 \texttt{Altitude} + \beta_2 \texttt{Temp} + \beta_3 \texttt{Cond} + \beta_4 \texttt{Pente} + \beta_5 \texttt{Larg} + \beta_6 \texttt{Prof} + \beta_7 \texttt{V.Fond} + \beta_8 \texttt{V.Surf} + \beta_9 \texttt{Abris} + \beta_{10} \texttt{DensInv} + \beta_{11} \texttt{Module} + \beta_{12} \texttt{DebitE} + \epsilon$$

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.3635609  4.0054570   3.336  0.00329 **
Altitude    -0.0031159  0.0013832  -2.253  0.03567 *
Temp        -0.2495276  0.2148390  -1.161  0.25913
Cond         0.0014675  0.0034465   0.426  0.67480
Pente       -0.0563558  0.0567191  -0.994  0.33229
Larg        -0.0890235  0.0527550  -1.687  0.10704
Prof        -1.7418561  1.9553569  -0.891  0.38362
V.Fond      -0.4704475  1.5234659  -0.309  0.76067
V.Surf      -1.1883832  0.5512356  -2.156  0.04345 *
Abris        0.0306894  0.0110323   2.782  0.01151 *
DensInv      0.0003378  0.0004689   0.721  0.47955
Module       0.0002368  0.0001217   1.946  0.06587 .
DebitE       0.0073534  0.0064769   1.135  0.26966
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4193 on 20 degrees of freedom
```

Conclure (grâce au test de Fisher) que la régression est, dans son ensemble, significative n'implique pas nécessairement que toutes les variables explicatives ont une contribution significative.

- Comment choisir les variables explicatives les plus pertinentes ?

Sélection des variables

(Quelques algorithmes "automatiques" d'aide à la décision)

On souhaite sélectionner parmi les p variables explicatives, celles qui donnent le "meilleur" modèle pour prédire la réponse Y.
On cherche donc :

- un critère de qualité d'un modèle afin de comparer deux modèles n'ayant pas nécessairement le même nombre de variables explicatives.
- une procédure qui permet de choisir parmi tous les modèles : On parle de procédure de choix de modèle.
- Comme la complexité augmente avec le nombre de variables : il y a $2^p - 1$ modèles possibles (avec $p = 12$: 4095 modèles possibles ...) En pratique, on utilise donc des heuristiques dont les plus simples sont les procédures pas à pas ascendante ou descendante.

Méthode 1 : élimination pas à pas à l'aide du test de Student

- **méthode "descendante"** : on estime le modèle complet avec les p variables disponibles. On retire du modèle la variable dont la p-value du test de Student est la moins significative (la plus grande). Et ainsi de suite ...jusqu'à ce que toutes les variables restantes soient "significatives" pour un seuil α donné.

Variables éliminées : `V.Fond`, `Cond`, `DensInv`, `Prof`, `Pente`, `Temp`, `DebitE`

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.278e+00  5.792e-01  16.017  2.6e-15 ***
Altitude    -1.918e-03  5.005e-04  -3.832  0.00069 ***
Larg        -1.150e-01  3.750e-02  -3.065  0.00489 **
V.Surf      -7.806e-01  3.507e-01  -2.226  0.03458 *
Abris        1.934e-02  5.829e-03   3.318  0.00260 **
Module       1.891e-04  8.445e-05   2.239  0.03356 *
---
Multiple R-squared:  0.7104,Adjusted R-squared:  0.6568
F-statistic: 13.25 on 5 and 27 DF,  p-value: 1.441e-06
```

Méthode 2 : élimination pas à pas à l'aide de critères : le $r^2_{ajusté}$

Les coefficients r^2 et $r^2_{ajusté}$:

- Le coefficient $r^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$:
  - mesure l'"ajustement" du modèle aux données,
  - augmente lorsque le nombre de variables incluses dans le modèle augmente,
  - permet de comparer des modèles ayant le même nombre de variables.
- Le coefficient $r^2_{ajusté} = 1 - \frac{SSR/(n-p-1)}{SST/(n-1)}$:
  - n'augmente pas forcément lorsque le nombre de variables introduites dans le modèle augmente,
  - permet de comparer des modèles ayant un nombre de variables différent.

Méthode 3 : élimination pas à pas à l'aide de critères pénalisés AIC ou BIC

On minimise ces critères dans une procédure de sélection de variables :

$$AIC = n \log(SSR) + 2 \times k \quad \text{(Akaike Information Criterion)}$$

et

$$BIC = n \log(SSR) + k \log(n) \quad \text{(Bayesian Information Criterion)}$$

Le critère BIC aura tendance à pénaliser davantage les modèles avec un grand nombre de variables.

Méthode 3 : élimination pas à pas **"descendante" AIC** :

- on estime le modèle complet avec les p variables disponibles. A chaque étape on calcule l'AIC en retirant du modèle une des variables, et on garde le modèle qui donne le critère *AIC* le plus petit. On arrête lorsque le critère *AIC* ne diminue plus lorsqu'on retire une variable de plus.

Variables éliminées (dans l'ordre) : `V.Fond`, `Cond`, `DensInv`, `Prof`

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.438e+01  3.412e+00   4.215 0.000306 ***
Altitude    -3.599e-03  1.125e-03  -3.200 0.003844 **
Temp        -2.858e-01  1.828e-01  -1.563 0.131076
Pente       -6.537e-02  4.976e-02  -1.314 0.201392
Larg        -1.032e-01  4.170e-02  -2.474 0.020831 *
Abris        2.329e-02  6.427e-03   3.624 0.001354 **
V.Surf      -1.306e+00  4.886e-01  -2.673 0.013319 *
Module       1.885e-04  8.188e-05   2.303 0.030284 *
DebitE       7.976e-03  5.014e-03   1.591 0.124770
---
AIC = -53.683
Multiple R-squared:  0.7738,Adjusted R-squared:  0.6984
F-statistic: 10.26 on 8 and 24 DF,  p-value: 4.049e-06
```

Méthode 3 : élimination pas à pas **"descendante" BIC** :

- on part du modèle complet avec les p variables disponibles. A chaque étape on calcule le BIC en retirant du modèle une des variables, et on garde le modèle qui donne le critère *BIC* le plus petit. On arrête lorsque le critère *BIC* ne diminue plus lorsqu'on retire une variable de plus.

Variables éliminées (dans l'ordre) : `V.Fond`, `Cond`, `DensInv`, `Prof`, `Pente`, `Temp`

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.226e+00  5.489e-01  16.808 1.75e-15 ***
Altitude    -2.205e-03  4.944e-04  -4.460  0.00014 ***
Larg        -1.343e-01  3.675e-02  -3.654  0.00115 **
Abris        2.009e-02  5.530e-03   3.632  0.00121 **
V.Surf      -7.366e-01  3.327e-01  -2.214  0.03580 *
Module       2.149e-04  8.094e-05   2.655  0.01335 *
DebitE       9.956e-03  4.898e-03   2.033  0.05240 .
---
BIC=-43,925
Multiple R-squared:  0.7501,Adjusted R-squared:  0.6925
F-statistic: 13.01 on 6 and 26 DF,  p-value: 9.168e-07
```

Comparaison de modèles "emboîtés"

Les 3 méthodes ne sélectionnent pas les mêmes variables.

– Model 0 : sélection par test de Student :

5 Variables sélectionnées : `Altitude`, `Larg`, `V.Surf`, `Abris`, `Module`

– Model 1 : sélection par BIC :

6 Variables sélectionnées : `Altitude`, `Larg`, `V.Surf`, `Abris`, `Module`, `DebitE`

– Model 2 : sélection par AIC : 8 Variables sélectionnées : `Altitude`, `Temp`, `Pente`, `Larg`, `V.Surf`, `Abris`, `Module`, `DebitE`

Les modèles sont "emboîtés" (les variables du "plus petit" modèle sont toutes présentes dans le modèle "plus grand") :
Model 0 ("student") est emboîté dans Model 1 (BIC) et dans Model 2 (AIC).
Model 1 (BIC) est emboîté dans Model 2 (AIC).

Comment les comparer ? avec des modèles emboîtés, on peut réaliser un test de Fisher (encore lui !)

Comparaison de modèles "emboîtés" : test de Fisher

On teste $\mathcal{H}_0$: Model 0 ("student") contre $\mathcal{H}_0$: Model 1 (BIC) :
Si on rejette $\mathcal{H}_0$ alors on garde les variables du Model 1.

Exemple : (sortie R)

```
Model 0: logBiomTot ~ Altitude + Larg + Abris + V.Surf + Module
Model 1: logBiomTot ~ Altitude + Larg + Abris + V.Surf + Module
                      + DebitE
  Res.Df    RSS    Df Sum of Sq      F     Pr(>F)
1     27   4.8128
2     26   4.1528  1   0.66003      4.1324 0.0524
```

Conclusion : au risque de 5%, on garde le Model 0 plutôt que le Model 1 (de justesse mais on ne rejette pas $\mathcal{H}_0$!).

Comparaison de modèles "emboîtés" : test de Fisher

On teste $\mathcal{H}_0$: Model 1 ("BIC") contre $\mathcal{H}_0$: Model 2 (AIC) :
Si on rejette $\mathcal{H}_0$ alors on garde les variables du Model 2, sinon il n'y a pas d'intérêt à ajouter les variables du Model 2 et donc on garde le Model 1.

Exemple : (sortie R)

```
Model 1:logBiomTot ~Altitude + Larg + Abris + V.Surf + Module + DebitE
Model 2:logBiomTot ~Altitude + Temp + Pente + Larg + Abris + V.Surf
                     + Module + DebitE
   Res.Df     RSS     Df   Sum of Sq      F    Pr(>F)
1      26     4.1528
2      24     3.7595  2    0.3933      1.2554  0.303
```

Conclusion : au risque de 5%, on garde le Model 1 plutôt que le Model 2 (on ne rejette pas $\mathcal{H}_0$).