Modèles linéaires et généralisés

Arnaud Poinas

Année 2024/2025

Table des matières

Introduction

Modalités du cours

Séances STDV :

- 10 séances de 2H de cours.
- 7 séances de 2H de TD.
- 6 séances de 2H et 1 séances de 4H de TP en R.

Évaluation STDV :

- 1 compte-rendu de TP (25%)
- 1 DM (25%)
- 1 examen terminal (50%)

⚠ Il n'y aura pas de seconde session.

Séances MFA :

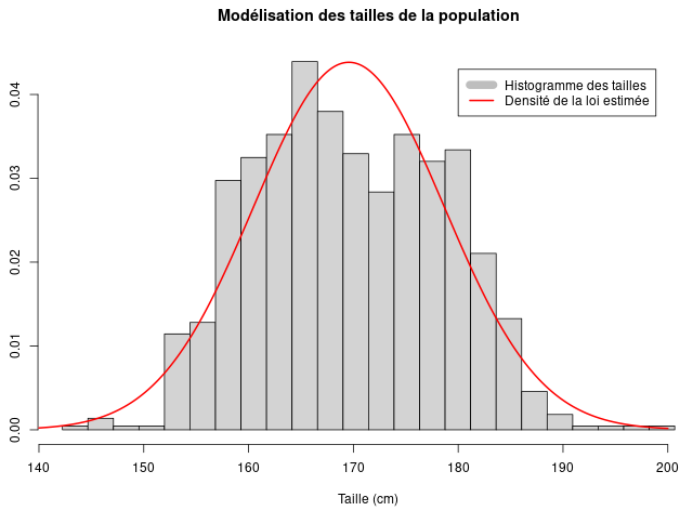
- 5 séances de 2H de cours.
- 4 séances de 2H de TD.
- 3 séances de 2H de TP en Python.

Évaluation MFA :

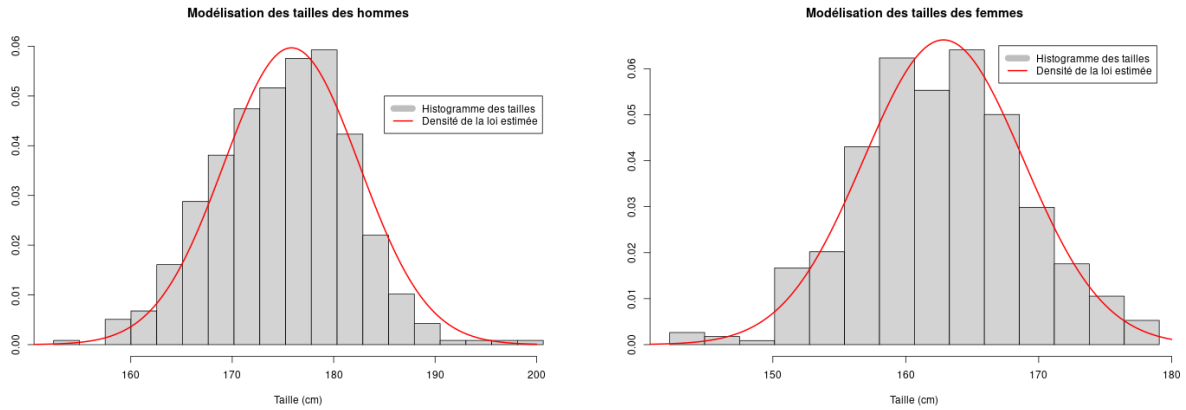
- 1 DM (50%)
- 1 examen terminal (50%)

Présentation du cours

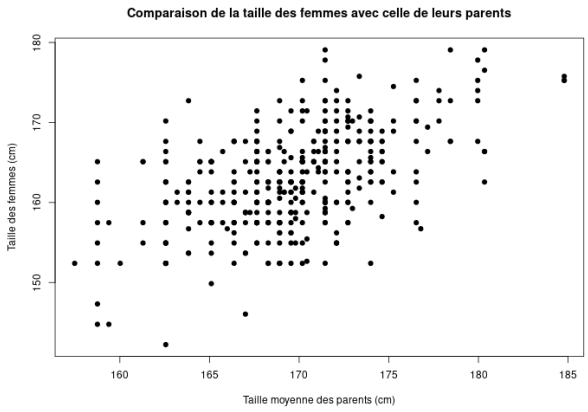
Ce cours fait suite au cours de statistiques inférentielles dans lequel on considérait avoir des données $y_1, \cdots, y_n$ issues de variables aléatoires $Y_1, \cdots, Y_n$ i.i.d. issues d'une famille paramétrique de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$. En général, un tel contexte est trop limité en pratique. Pour illustrer ces limites, on considère le célèbre jeu de données obtenu par Sir Francis Galton (années $1880-1890$) contenant la taille et le sexe de 898 adultes ainsi que la taille de leurs parents. Si on modélise la taille des individus par une loi normale et qu'on estime les paramètres associés alors on obtient l'histogramme suivant.



On remarque que le modèle colle assez bien aux données mais est loi d'être parfait. Ce modèle indique alors qu'un individu typique a 95% de chance d'avoir une taille dans l'intervalle $[151.7\text{cm}, 187.4\text{cm}]$. Par contre, si on sépare la population selon le sexe et qu'on modélise par une loi normale la taille des hommes et des femmes alors, après estimation des paramètres, on obtient les histogrammes suivants



Non seulement on peut voir que la modélisation colle mieux aux données mais en plus, on obtient des intervalles de confiances plus précis avec 95% de chance pour les hommes d'avoir une taille dans l'intervalle $[162.7\text{cm}, 188.9\text{cm}]$ et 95% de chance pour les femmes d'avoir une taille dans l'intervalle $[151.0\text{cm}, 174.6\text{cm}]$. On voit donc qu'on peut améliorer le modèle si on ne fait pas l'hypothèse que toutes les données sont issues de lois avec les mêmes paramètres mais en supposant à la place que les paramètres dépendent de quantités extérieures (ici le sexe). En suivant ce raisonnement, il est raisonnable de considérer que la taille d'un individu va être affectée par celle de ces parents. On a donc aussi envie de faire un modèle tel que la taille d'une personne suit une loi normale dont les paramètres dépendent de la taille des parents. Si on se concentre sur les données des femmes et qu'on regarde le nuage de point des tailles des femmes en fonction de la taille moyenne de leurs parents



alors on remarque une tendance linéaire entre ces deux quantités. Cela laisse suggérer la possibilité de modéliser la taille y_i des femmes dont les parents ont une taille moyenne x_i par une loi normale Y_i dont la moyenne possède une relation linéaire avec les x_i soit

$$Y_i \sim \mathcal{N}(\alpha x_i + \beta, \sigma^2)$$

où, de façon équivalente,

$$Y_i = \alpha x_i + \beta + \varepsilon_i \text{ avec } \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

C'est ce qu'on appelle **le modèle linéaire simple**. L'objectif du cours va être d'étudier les propriétés d'un tel modèle ainsi que de généraliser le principe avec d'autres lois et des relations plus compliquées.

Notations du cours

On considère un jeu de données de n individus et $p+1$ variables avec $p \geqslant 1$. L'objectif du cours est de modéliser le comportement de l'une des variables, appelée la **variable d'intérêt** (ou **variable expliquée**), en fonction des autres variables appelées les **variables explicatives**. On notera par la suite

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix},$$

le vecteur des valeurs de la variable d'intérêt et la matrice des variables explicatives. On suppose que les y_i sont issues de variables **indépendantes** Y_i dont la loi dépend des variables explicatives $x_{i,1}, \cdots, x_{i,p}$. Afin de simplifier un peu les notations on notera aussi

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

le vecteur aléatoires des Y_i. En général ce sera évident si Y réfère au vecteur des données ou le vecteur aléatoire.

Notations : Dans ce cours, pour des vecteurs X et Y de même taille on note avec des lettres minuscules $\mathrm{cov}(X,Y)$, $\mathrm{var}(X)$, et $\mathrm{corr}(X,Y)$ leur covariance, variance et corrélations. Pour des variables aléatoires réelles (v.a.r.) X et Y on note avec des premières lettres majuscules $\mathrm{Cov}(X,Y)$, $\mathrm{Var}(Y)$, et $\mathrm{Corr}(Y,Y)$ leur covariance, variance et corrélation. Par exemple, pour un vecteur $Y = (y_1, \cdots, y_n)$ et une v.a.r. Y on a

$$\mathrm{var}(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 \quad \text{et} \quad \mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2.$$

On note aussi $\bar{Y}$ pour la moyenne d'un vecteur Y et $\overline{Y^2}$ pour la moyenne des valeurs de Y au carré. On peut donc écrire par exemple $\mathrm{var}(Y) = \overline{Y^2} - \bar{Y}^2$.

Chapitre 1

Modèles linéaires

I Rappel sur les vecteurs gaussiens

On commence par rappeler succinctement les propriétés de base des vecteurs gaussiens que l'on utilisera dans la suite du cours.

Définition 1

Soit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ *un vecteur aléatoire à valeurs dans* $\mathbb{R}^d$. *X est dit* ***vecteur gaussien*** *si toute combinaison linéaire de ses coordonnées est une variable aléatoire réelle gaussienne (de variance éventuellement nulle).*
On note le vecteur espérance de X par

$$\mu = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix} \in \mathbb{R}^d$$

et la matrice de variance-covariance de X par

$$\Sigma = (Cov(X_i, X_j))_{1 \leqslant i,j \leqslant d} \in S_n^+(\mathbb{R}).$$

La loi de X est notée $\mathcal{N}_d(\mu, \Sigma)$.

Proposition 2

- *Soit* $X \sim \mathcal{N}_d(\mu, \Sigma)$ *un vecteur gaussien dans* $\mathbb{R}^d$. *Soient* $A \in \mathcal{M}_{k,d}(\mathbb{R})$ *et* $b \in \mathbb{R}^k$. *Alors*

$$AX + b \sim \mathcal{N}_k(A\mu + b, A\Sigma\ {}^tA).$$

- *Soit* $X \sim \mathcal{N}_d(\mu, \Sigma)$. *Les coordonnées de X sont indépendantes ssi* Σ *est diagonale.*
- *Soit* $X \sim \mathcal{N}_d(\mu, \Sigma)$. *Si* $\det \Sigma \neq 0$, *alors X admet la densité suivante sur* $\mathbb{R}^d$:

$$\forall x \in \mathbb{R}^d,\ f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle}.$$

Théorème 3 (*Cochran*)

Soit $\sigma^2 \in \mathbb{R}_+^*$ *et* $X \sim \mathcal{N}_d(0_d, \sigma^2 I_d)$. *Soient* $E_1, \cdots, E_k$ *des sous-espaces vectoriels orthogo-*

naux de $\mathbb{R}^d$. On note P_{E_i} la projection sur l'espace E_i. Alors, les vecteurs $P_{E_i}(X)$ sont des vecteurs gaussiens indépendants vérifiant

$$\frac{1}{\sigma^2}\|P_{E_i}(X)\|_2^2 \sim \chi^2(\dim(E_i)).$$

Démonstration : Soit E_{k+1} le complément orthogonal de $E_1 \oplus \cdots \oplus E_k$. On pose $r_i = \dim(E_i)$ et $e_{i,1}, \cdots, e_{i,r_i} \in \mathbb{R}^d$ une base orthonormale de E_i. Comme les E_i sont orthogonaux et $\mathbb{R}^d = E_1 \oplus \cdots \oplus E_{k+1}$ on en déduit que $e_{1,1}, \cdots, e_{1,r_1}, \cdots, e_{k+1,1}, \cdots, e_{k+1,r_{k+1}}$ est une base orthonormale de $\mathbb{R}^d$. On pose P_i la matrice de taille $n \times r_i$ dont les lignes sont $e_{i,1}, \cdots, e_{i,r_i}$. Ces matrices vérifient dont $P_i P_i^T = I_{r_i}$ et $P_{E_i}(X) = P_i^T P_i X$. Maintenant on pose P la matrice orthogonale qui contient tout les $e_{i,j}$:

$$P = \begin{pmatrix} P_1 \\ \vdots \\ P_{k+1} \end{pmatrix} = \begin{pmatrix} e_{1,1} \\ \vdots \\ e_{1,r_1} \\ \vdots \\ e_{k+1,1} \\ \vdots \\ e_{k+1,r_{k+1}} \end{pmatrix} \Rightarrow PX = \begin{pmatrix} P_1 X \\ \vdots \\ P_{k+1} X \end{pmatrix}$$

Comme P est orthogonale alors $PP^T = I_n$ donc $PX \sim \mathcal{N}_d(P0_d, \sigma^2 PP^T) = \mathcal{N}_d(0_d, \sigma^2 I_d)$. On en déduit alors que les coordonnées de PX (et donc les $P_i X$) sont indépendants d'où les $P_i^T P_i X = P_{E_i}(X)$ sont indépendants. De plus, on a

$$P_i X \sim \mathcal{N}_d(0_d, \sigma^2 I_{r_i}) \Rightarrow \frac{1}{\sigma} P_i X \sim \mathcal{N}_d(0_d, I_{r_i}).$$

Les coordonnées de $\frac{1}{\sigma} P_i X$ sont donc des variables i.i.d. de loi $\mathcal{N}(0,1)$ donc la somme de leur carrées suit une loi $\chi^2(r_i)$ d'où

$$\frac{1}{\sigma^2}\|P_i X\|_2^2 \sim \chi^2(r_i)$$

On conclue alors en remarquant :

$$\|P_{E_i}(X)\|_2^2 = \langle P_i^T P_i X, P_i^T P_i X \rangle = \langle P_i X, P_i P_i^T P_i X \rangle = \langle P_i X, P_i X \rangle = \|P_i X\|_2^2.$$

■

II Modèle linéaire simple

On commencer par illustrer le principe et les résultat de base des modèles linéaire en considérant le cas (non-trivial) le plus simple : le modèle linéaire simple. Néanmoins, tous les résultats de cette partie seront généralisés dans le section suivante.

1 Écriture du modèle

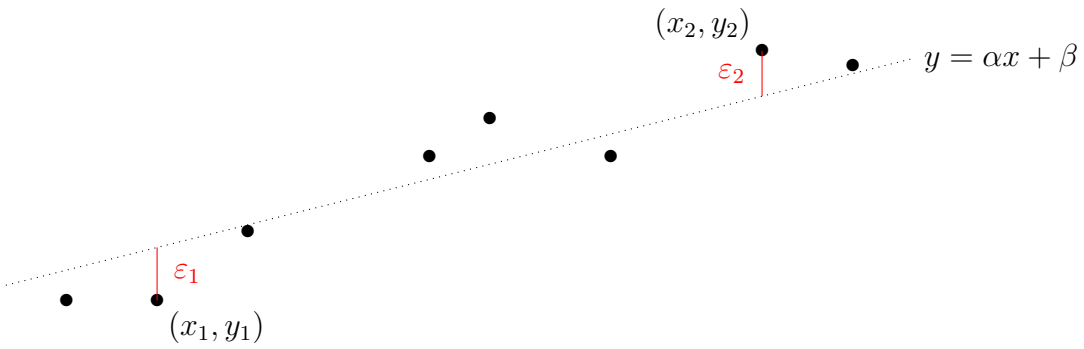
On considère qu'on a une variable d'intérêt quantitative et une seule variable explicative quantitative d'où $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$. De plus, on suppose aussi que $\text{var}(X) > 0$. Le **modèle linéaire** consiste à supposer que les y_i sont issus de v.a.r. Y_i vérifiant

$$Y_i = \alpha x_i + \beta + \varepsilon_i, \text{ où } \varepsilon_i \text{ i.i.d. de loi } \mathcal{N}(0, \sigma^2).$$

Autrement dit, on considère que $Y_i \sim \mathcal{N}(\alpha x_i + \beta, \sigma^2)$. On peut réécrire ce modèle vectoriellement sous la forme

$$Y = \alpha X + \beta 1_n + \varepsilon \text{ où } 1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathcal{N}_n(0_n, \sigma^2 I_n).$$

On appelle **régression linéaire** le fait d'**ajuster** un tel modèle à un jeu de données, c'est à dire d'estimer α, β et σ^2.



Définition 4

*Soient $\hat{\alpha}$ et $\hat{\beta}$ des estimateurs de α et β. Pour tout i on note $\hat{y}_i = \hat{\alpha} x_i + \hat{b}$ la valeur de y_i prédite par le modèle. La quantité $y_i - \hat{y}_i$ est alors appelée le i-ème **résidu** du modèle. On note $\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$ le vecteur des prédictions et $Y - \hat{Y}$ le vecteur des résidus.*

Remarque: Si on observe une nouvelle valeur x_{n+1} de la variable explicative alors on peut estimer la valeur de la variable d'intérêt associée par $\hat{y}_{n+1} = \hat{\alpha} x_{n+1} + \hat{\beta}$ même si elle n'est pas connue. L'utilisation d'un modèle permet donc de faire de la prédiction.

2 Estimation des paramètres

On souhaite estimer les paramètres α, β et σ^2 par maximum de vraisemblance. Comme les y_i sont issues de variables indépendantes Y_i alors on a

$$L(\alpha, \beta, \sigma^2 | y_1, \cdots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i),$$

où f_{Y_i} est la densité de $\mathcal{N}(\alpha x_i + \beta, \sigma^2)$, la loi de Y_i, d'où

$$L(\alpha, \beta, \sigma^2 | y_1, \cdots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha x_i - \beta)^2}$$

Théorème 5

Les estimateurs du maximum de vraisemblance de α, β et σ^2 existent, sont uniques et vérifient :

$$\hat{\alpha} = \frac{cov(X, Y)}{var(X)}, \ \hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} \text{ et } \hat{\sigma}^2 = \frac{1}{n}\|Y - \hat{Y}\|_2^2.$$

Démonstration : La log-vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}(\alpha,\beta,\sigma^2) &= \sum_{i=1}^{n}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i-\alpha x_i-\beta)^2\right)\\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha x_i-\beta)^2.\end{aligned}$$

On remarque que quelque soit la valeur de σ^2, maximiser $\mathcal{L}$ par rapport à α et β revient à minimiser la quantité $\sum_{i=1}^n(y_i-\alpha x_i-\beta)^2$ qui ne dépend pas de σ^2. Si on note $\hat\alpha,\hat\beta$ et $\hat\sigma^2$ des paramètres minimisant $\mathcal{L}$ alors $(\hat\alpha,\hat\beta)$ minimise la fonction $g(\alpha,\beta)=\sum_{i=1}^n(y_i-\alpha x_i-\beta)^2$ et σ^2 maximise la fonction

$$\phi : s\mapsto \mathcal{L}(\hat\alpha,\hat\beta,s) = -\frac{n}{2}\log(2\pi s) - \frac{1}{2s}\sum_{i=1}^{n}(y_i-\hat\alpha x_i-\hat\beta)^2 = -\frac{n}{2}\log(2\pi s) - \frac{1}{2s}\|Y-\hat Y\|_2^2.$$

Calcul de $\hat\alpha$ et $\hat\beta$:
On commence par calculer le gradient et la Hessienne de la fonction $g(\alpha,\beta)$.

$$\nabla g(\alpha,\beta) = \begin{pmatrix}\frac{\partial g}{\partial\alpha}\\ \frac{\partial g}{\partial\beta}\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n -2x_i(y_i-\alpha x_i-\beta)\\ \sum_{i=1}^n -2(y_i-\alpha x_i-\beta)\end{pmatrix};$$

$$H(g)(\alpha,\beta) = \begin{pmatrix}\frac{\partial^2 g}{\partial\alpha^2} & \frac{\partial^2 g}{\partial\alpha\partial\beta}\\ \frac{\partial^2 g}{\partial\beta\partial\alpha} & \frac{\partial^2 g}{\partial\beta^2}\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n 2x_i^2 & \sum_{i=1}^n 2x_i\\ \sum_{i=1}^n 2x_i & 2n\end{pmatrix}.$$

On commence par remarque que la Hessienne est tout le temps une matrice symétrique définie positive car $\mathrm{Tr}(H(g)(\alpha,\beta)) = \sum_{i=1}^n 2x_i^2 + 2n \geqslant 0$ et

$$\det(H(g)(\alpha,\beta)) = 4n\sum_{i=1}^n x_i^2 - 4\left(\sum_{i=1}^n x_i\right)^2 = 4n^2\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right) = 4n^2\mathrm{var}(X) > 0.$$

Du coup, g est une fonction strictement convexe et donc tout (α,β) qui annule ∇g est un minimum global. Or,

$$\begin{aligned}\nabla g(\hat\alpha,\hat\beta) = 0 &\Leftrightarrow \begin{cases}\sum_{i=1}^n -2x_i(y_i-\hat\alpha x_i-\hat\beta) = 0\\ \sum_{i=1}^n -2(y_i-\hat\alpha x_i-\hat\beta) = 0\end{cases}\\ &\Leftrightarrow \begin{cases}\hat\alpha\sum_{i=1}^n x_i^2 + \hat\beta\sum_{i=1}^n x_i = \sum_{i=1}^n x_iy_i\\ \hat\alpha\sum_{i=1}^n x_i + \hat\beta n = \sum_{i=1}^n y_i\end{cases}\\ &\Leftrightarrow \begin{pmatrix}\sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & n\end{pmatrix}\begin{pmatrix}\hat\alpha\\ \hat\beta\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n x_iy_i\\ \sum_{i=1}^n y_i\end{pmatrix}\end{aligned}$$

d'où

$$\hat\alpha = \frac{n\sum_{i=1}^n x_iy_i - \sum_{i=1}^n x_i\sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^n x_iy_i - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

et

$$\sum_{i=1}^n -2(y_i-\hat\alpha x_i-\hat\beta) = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^n(y_i-\hat\alpha x_i-\hat\beta) = 0 \Leftrightarrow \hat\beta = \bar Y - \hat\alpha\bar X.$$

Calcul de $\hat\sigma^2$:
On a

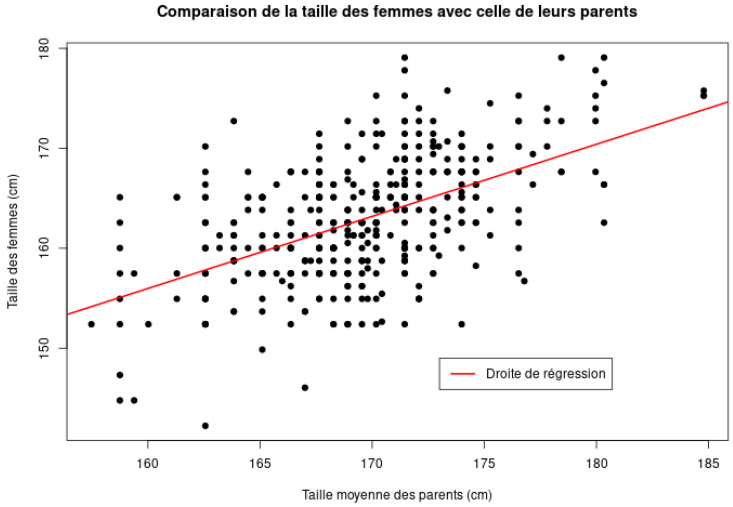
$$\phi'(s) = -\frac{n}{2s} + \frac{1}{2s^2}\|Y-\hat Y\|_2^2 = \frac{n}{2s^2}\left(s - \frac{1}{n}\|Y-\hat Y\|_2^2\right).$$

La fonction ϕ possède donc le tableau de variations suivant :

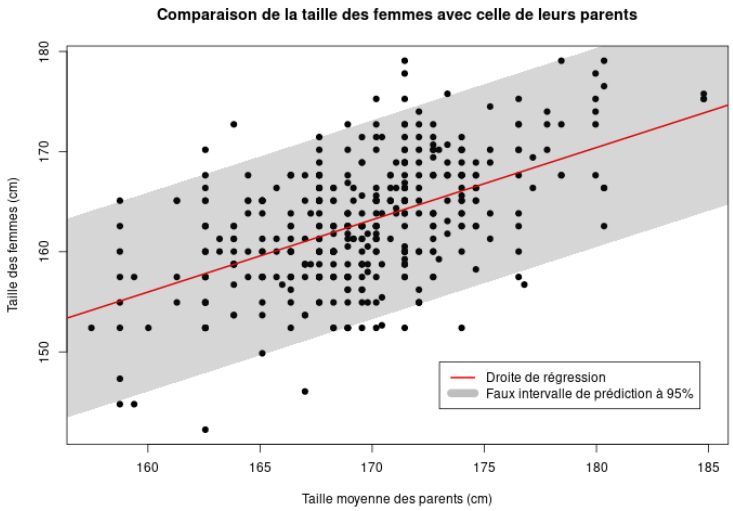
s	0	$\frac{1}{n}\|Y-\hat{Y}\|_2^2$	$+\infty$
$\phi'(s)$	>0	0	<0
$\phi(s)$	↗		↘

On en déduit que ϕ a un unique maximum en $\hat{\sigma}^2 = \frac{1}{n}\|Y-\hat{Y}\|_2^2$. ■

Exemple: Si on estime les paramètres α, β et σ sur les données de taille de Sir Francis Galton et que l'on trace la droite $y = \hat{\alpha}x + \hat{\beta}$, appelée **droite de régression**, on obtient le résultat suivant qui correspond bien aux données.

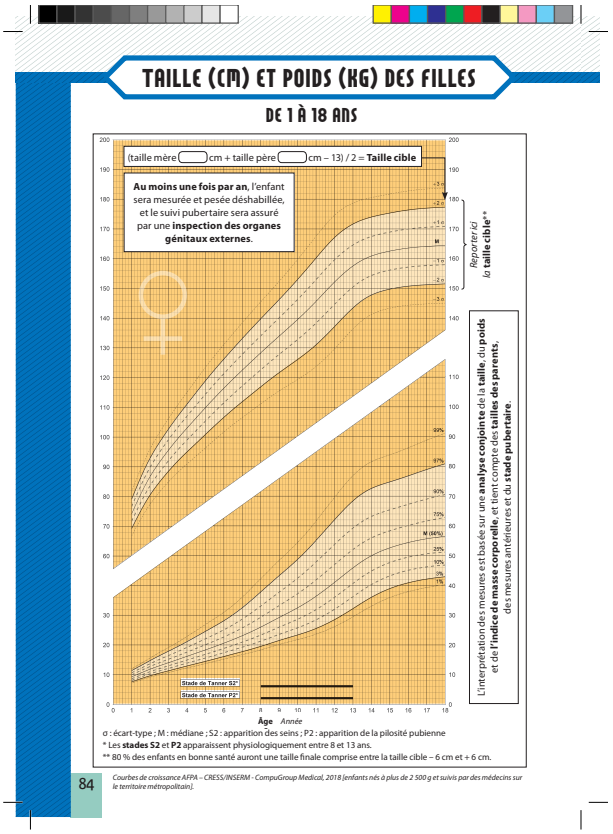


De plus, si on fait comme si les paramètres $\hat{\alpha}$, $\hat{\beta}$ et $\hat{\sigma}$ sont les vrai paramètres, alors pour une valeur x de la variable explicative il y a 95% de chance que la variable d'intérêt possède une valeur dans l'intervalle $[\hat{\alpha}x+\hat{\beta}-1.96\hat{\sigma}, \hat{\alpha}x+\hat{\beta}+1.96\hat{\sigma}]$. Si on rajoute ces deux droites au nuage de points on observe bien que la majorité des données se trouvent dans cet intervalle.

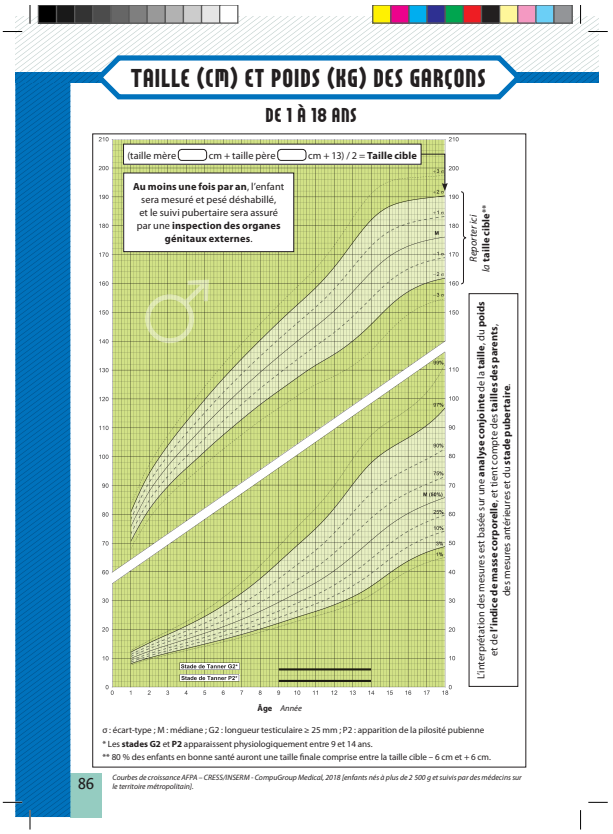


On verra plus tard comment faire un vrai intervalle de confiance pour prédire les valeurs de la variable d'intérêt.

Remarque: Ce genre de modèle est utilisé pour construire les courbes de croissances où on modélise la taille d'un enfant par une loi normale dont l'espérance et la variance dépend du sexe de l'enfant et de son age. Cela permet notamment de détecter les enfants possédant une croissance anormale.



(a) Courbes de taille et poids des filles



(b) Courbes de taille et poids des garçons

A noter que nos données concernent la taille à l'âge adulte donc nous n'aurons pas à regarder l'effet de l'âge. On remarquera aussi sur ces courbes que seules les données de taille sont supposées issues d'une loi normale. Ce n'est pas le cas des données de poids.

3 Loi des estimateurs

Proposition 6

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n var(X)}\right) \quad et \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2 \overline{X^2}}{n var(X)}\right).$$

En particulier, $\hat{\alpha}$ et $\hat{\beta}$ sont des estimateurs non biaisés. De plus,

$$cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \overline{X}}{n var(X)}.$$

Démonstration : $\hat{\alpha}$ et $\hat{\beta}$ suivent des lois normales car ils s'écrivent comme combinaison linéaire

des $Y_i \sim \mathcal{N}(\alpha x_i + \beta, \sigma^2)$ qui sont indépendantes. On commence par calculer leur espérance :

$$\begin{aligned}
\mathbb{E}[\hat{\alpha}] &= \frac{\frac{1}{n}\sum_{i=1}^n x_i \mathbb{E}[Y_i] - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i]\right)}{\mathrm{var}(X)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n x_i(\alpha x_i + \beta) - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n (\alpha x_i + \beta)\right)}{\mathrm{var}(X)} \\
&= \frac{\frac{\alpha}{n}\sum_{i=1}^n x_i^2 + \frac{\beta}{n}\sum_{i=1}^n x_i - \alpha\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 - \frac{\beta}{n}\sum_{i=1}^n x_i}{\mathrm{var}(X)} = \alpha
\end{aligned}$$

et

$$\mathbb{E}[\hat{\beta}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] - \mathbb{E}[\hat{\alpha}]\bar{X} = \frac{1}{n}\sum_{i=1}^n (\alpha x_i + \beta) - \alpha\bar{X} = \beta.$$

Pour calculer la variance de α, il faut déjà commencer à l'écrire comme une somme de variables indépendantes

$$\hat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^n x_i Y_i - \bar{X} \times \frac{1}{n}\sum_{i=1}^n Y_i}{\mathrm{var}(X)} = \frac{1}{n\mathrm{var}(X)}\sum_{i=1}^n (x_i - \bar{X})Y_i.$$

On en déduit

$$\begin{aligned}
\mathrm{Var}(\hat{\alpha}) &= \frac{1}{n^2\mathrm{var}(X)^2}\sum_{i=1}^n \mathrm{var}((x_i - \bar{X})Y_i) \\
&= \frac{1}{n^2\mathrm{var}(X)^2}\sum_{i=1}^n (x_i - \bar{X})^2\sigma^2 \\
&= \frac{n\mathrm{var}(X)}{n^2\mathrm{var}(X)^2}\sigma^2 \\
&= \frac{\sigma^2}{n\mathrm{var}(X)}.
\end{aligned}$$

On fait de même pour β. On commence par l'écrire comme une somme de variables indépendantes :

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^n Y_i - \frac{\bar{X}}{n\mathrm{var}(X)}\sum_{i=1}^n (x_i - \bar{X})Y_i = \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\bar{X}(x_i - \bar{X})}{\mathrm{var}(X)}\right) Y_i$$

d'où on en déduit

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \frac{1}{n^2}\sum_{i=1}^n \left(1 - \frac{\bar{X}(x_i - \bar{X})}{\mathrm{var}(X)}\right)^2 \sigma^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n \left(1 - \frac{2\bar{X}(x_i - \bar{X})}{\mathrm{var}(X)} + \frac{\bar{X}^2(x_i - \bar{X})^2}{\mathrm{var}(X)^2}\right)\sigma^2 \\
&= \frac{\sigma^2}{n^2}\left(n + \frac{n\bar{X}^2}{\mathrm{var}(X)}\right) \\
&= \frac{\overline{X^2}\sigma^2}{n\mathrm{var}(X)}.
\end{aligned}$$

Ensuite, on a

$$\mathrm{cov}(\hat{\alpha}, \hat{\beta}) = \mathrm{cov}\left(\frac{1}{n\mathrm{var}(X)}\sum_{i=1}^n (x_i - \bar{X})Y_i,\ \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\bar{X}(x_i - \bar{X})}{\mathrm{var}(X)}\right) Y_i\right).$$

Comme $\text{cov}(Y_i, Y_j) = 0$ si $i \neq j$ et $\text{cov}(Y_i, Y_i) = \sigma^2$ sinon alors on en déduit :

$$\begin{aligned}
\text{cov}(\hat{\alpha}, \hat{\beta}) &= \frac{1}{n^2 \text{var}(X)} \sum_{i=1}^{n} (x_i - \bar{X}) \left(1 - \frac{\bar{X}(x_i - \bar{X})}{\text{var}(X)}\right) \sigma^2 \\
&= \frac{\sigma^2}{n^2 \text{var}(X)} \left(\sum_{i=1}^{n} (x_i - \bar{X}) - \sum_{i=1}^{n} \frac{\bar{X}(x_i - \bar{X})^2}{\text{var}(X)}\right) \\
&= \frac{\sigma^2}{n^2 \text{var}(X)} \left(0 - \frac{\bar{X}\text{var}(X)}{\text{var}(X)}\right) \\
&= -\frac{\sigma^2 \bar{X}}{n \text{var}(X)}.
\end{aligned}$$

■

Proposition 7

Si $n \geqslant 2$ alors

$$n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

De plus, $\hat{\sigma}^2$ est indépendant de $\hat{\alpha}$ et $\hat{\beta}$.

Démonstration : On fait l'hypothèse que $\bar{X} = 0 = \langle X, 1_n \rangle$ pour simplifier les calculs. En utilisant l'écriture vectorielle du modèle linéaire, on a $Y = \alpha X + \beta 1_n + \varepsilon$ et $\hat{Y} = \hat{\alpha} X + \hat{\beta} 1_n$. On en déduit

$$\hat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \frac{\langle X, Y \rangle}{\langle X, X \rangle} = \frac{\langle X, \alpha X + \beta 1_n + \varepsilon \rangle}{\langle X, X \rangle} = \alpha + \frac{\langle X, \varepsilon \rangle}{\langle X, X \rangle}$$

$$\text{et } \hat{\beta} = \bar{Y} = \frac{1}{n}\langle 1_n, Y \rangle = \frac{1}{n}\langle 1_n, \alpha X + \beta 1_n + \varepsilon \rangle = \beta + \frac{1}{n}\langle 1_n, \varepsilon \rangle$$

donc

$$Y - \hat{Y} = (\alpha - \hat{\alpha})X + (\beta - \hat{\beta})1_n + \varepsilon = \varepsilon - \frac{\langle X, \varepsilon \rangle}{\langle X, X \rangle} X - \frac{1}{n}\langle 1_n, \varepsilon \rangle 1_n.$$

Le deuxième terme de $Y - \hat{Y}$ correspond à la projection orthogonale de ε sur l'espace engendré par X et le troisième terme correspond à la projection orthogonale de ε sur l'espace engendré par 1_n. Comme $\langle X, 1_n \rangle = 0$ alors ces espaces sont orthogonaux. $Y - \hat{Y}$ est donc la projection orthogonale de ε sur l'espace complémentaire de $\text{Vect}(1_n, X)$ qui est de dimension $n - 2$. Par le théorème de Cochran, on en déduit donc que

$$n\frac{1}{\sigma^2}\|Y - \hat{Y}\|^2 \sim \chi^2(n-2).$$

De plus, le théorème de Cochran nous dit que $Y - \hat{Y}$ est indépendant de la projection de ε sur X, c'est à dire $\frac{\langle X, \varepsilon \rangle}{\langle X, X \rangle} X = (\hat{\alpha} - \alpha)X$, et la projection de ε sur 1_n, c'est à dire $\frac{1}{n}\langle 1_n, \varepsilon \rangle 1_n = (\hat{\beta} - \beta)1_n$. Cela prouve alors l'indépendance de $\hat{\sigma}^2$ avec $\hat{\alpha}$ et $\hat{\beta}$. ■

Corollaire 8

On a $\mathbb{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$ donc $\hat{\sigma}^2$ est un estimateur biaisé de la variance. On définit alors l'estimateur

$$\tilde{\sigma}^2 := \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\|Y - \hat{Y}\|_2^2$$

qui est non biaisé et vérifie

$$(n-2)\frac{\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Remarque: En pratique on utilise plus souvent cet estimateur de σ^2 au lieu de l'estimateur du maximum de vraisemblance.

Une autre conséquence de la loi de $\hat{\sigma}^2$ est la possibilité d'exprimer la loi de $\hat{\alpha}-\alpha$ et $\hat{\beta}-\beta$ mais sans utiliser σ^2.

Proposition 9

$$\sqrt{n\text{var}(X)}\frac{\hat{\alpha}-\alpha}{\widetilde{\sigma}} \sim \mathcal{T}(n-2) \quad et \quad \sqrt{\frac{n\text{var}(X)}{\overline{X^2}}}\frac{\hat{\beta}-\beta}{\widetilde{\sigma}} \sim \mathcal{T}(n-2).$$

Démonstration : Rappel : Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2(n)$ avec $X \perp\!\!\!\perp Y$ alors $\frac{X}{\sqrt{Y/n}} \sim \mathcal{T}(n)$.

On a vu que $\sqrt{n\text{var}(X)}\frac{\hat{\alpha}-\alpha}{\sigma} \sim \mathcal{N}(0,1)$ et $(n-2)\frac{\widetilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ donc

$$\frac{\sqrt{n\text{var}(X)}\frac{\hat{\alpha}-\alpha}{\sigma}}{\sqrt{\frac{(n-2)\widetilde{\sigma}^2}{(n-2)\sigma^2}}} \sim \mathcal{T}(n-2) \Leftrightarrow \sqrt{n\text{var}(X)}\frac{\hat{\alpha}-\alpha}{\widetilde{\sigma}} \sim \mathcal{T}(n-2).$$

Même raisonnement pour β. ■

4 Intervalle de confiance et test statistiques pour α et β

On note $t_q^{(n)}$ le quantile d'ordre q de la loi $\mathcal{T}(n)$. On obtient les résultats suivants comme conséquence directe des propriétés vues à la section précédente :

Proposition 10

Soit $q \in]0,1[$, alors les intervalles de confiances suivants :

$$IC_{1-q}(\alpha) = \left[\hat{\alpha} \pm t_{1-q/2}^{(n-2)}\sqrt{\frac{\widetilde{\sigma}^2}{n\text{var}(X)}}\right]; \quad IC_{1-q}(\beta) = \left[\hat{\beta} \pm t_{1-q/2}^{(n-2)}\sqrt{\frac{\widetilde{\sigma}^2\overline{X^2}}{n\text{var}(X)}}\right];$$

sont des intervalles de confiance exacts de niveau $1-q$ pour les paramètres α et β.

Pour les modèles linéaires, il est important de savoir si les paramètres α ou β sont significativement différent de 0 car si ce n'est pas le cas alors il est possible de les supposer nuls et de simplifier le modèle. On considère alors les deux tests suivants :

$$\mathcal{H}_0 : \alpha = 0 \text{ contre } \mathcal{H}_1 : \alpha \neq 0 \text{ et } \mathcal{H}_0' : \beta = 0 \text{ contre } \mathcal{H}_1' : \beta \neq 0.$$

On définit les statistiques de test suivantes :

$$T = \hat{\alpha}\sqrt{\frac{n\text{var}(X)}{\widetilde{\sigma}^2}} \text{ et } T' = \hat{\beta}\sqrt{\frac{n\text{var}(X)}{\widetilde{\sigma}^2\overline{X^2}}},$$

qui sont de loi $\mathcal{T}(n-2)$ sous $\mathcal{H}_0$ (resp. $\mathcal{H}_0'$) mais qui vont prendre des valeurs éloignées de 0 sous $\mathcal{H}_1$ (resp. $\mathcal{H}_1'$). On rejette alors l'hypothèse $\mathcal{H}_0$ au risque q si $|T| \geqslant t_{1-q/2}^{(n-2)}$ et on rejette l'hypothèse $\mathcal{H}_0'$ au risque q si $|T'| \geqslant t_{1-q/2}^{(n-2)}$

Exemple: Pour les données de taille de Sir Francis Galton on a $n = 433$, $T = 13.369$ et $T' = 4.432$. Comme $t_{0.975}^{(431)} \approx 1.97$ alors on rejette au risque 5% l'hypothèse que $\alpha = 0$ et l'hypothèse que $\beta = 0$. Ce modèle trouve donc qu'il y a un effet significatif de la taille moyenne des parents sur la taille des femmes.

III Modèle linéaire multiple

1 Écriture du modèle

On considère toujours qu'on a une variable d'intérêt quantitative mais on suppose maintenant que l'on possède p variables explicatives quantitatives. Le modèle linéaire multiple consiste à modéliser la loi des variables aléatoires Y_i qui ont engendrées les données y_i par :

$$Y_i = \theta_0 + \theta_1 x_{i,1} + \cdots + \theta_p x_{i,p} + \varepsilon_i, \text{ où } \varepsilon_i \text{ i.i.d. de loi } \mathcal{N}(0, \sigma^2).$$

Autrement dit, on considère que $Y_i \sim \mathcal{N}(\theta_0 + \sum_{k=1}^{p} \theta_k x_{i,k}, \sigma^2)$.

Remarques:

- A chaque variable explicative est associé un coefficient. On parle souvent de θ_i comme le coefficient du modèle associé à la i-ème variable.
- Si on modifie la valeur de $x_{i,j}$ en y ajoutant une constante C, sans toucher aux autres variables, alors cela revient à ajouter $C\theta_i$ à $\mathbb{E}[Y_i]$. **Un effet additif sur les variables explicatives à un effet additif sur la variable d'intérêt**. Il en vient l'interprétation suivante des θ_i :
  - Si $\theta_i = 0$ alors la variable explicative associée au coefficient n'a aucun effet sur la variable d'intérêt lorsque les autres variables explicatives sont fixées.
  - Si $\theta_i > 0$ alors une augmentation de la variable explicative associée au coefficient, lorsque les autres variables explicatives sont fixées, va entraîner une augmentation de la moyenne de la variable d'intérêt (et inversement).
  - Si $\theta_i < 0$ alors une augmentation de la variable explicative associée au coefficient, lorsque les autres variables explicatives sont fixées, va entraîner une diminution de la moyenne de la variable d'intérêt (et inversement).

Définition 11

*Le paramètre θ_0 est appelé **l'intercept**.*

Afin de réécrire le modèle vectoriellement on définit la quantité suivante :

Définition 12

On note X_e la matrice des variables explicatives X à laquelle on rajoute une première colonne composée de 1 :

$$X_e = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}.$$

On peut donc réécrire le modèle sous la forme $Y = X_e\theta + \varepsilon$ où $\theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_p \end{pmatrix}$ est le vecteur des paramètres et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ est un vecteur Gaussien de loi $\mathcal{N}_n(0_n, \sigma^2 I_n)$. Le vecteur des y_i est donc la simulation d'un vecteur Gaussien de loi $\mathcal{N}_n(X_e\theta, \sigma^2 I_n)$. Les paramètres du modèle sont $\theta \in \mathbb{R}^{p+1}$ et $\sigma > 0$.

Remarque: On peut aussi considérer un modèle sans intercept ($\theta_0 = 0$). Dans ce cas-là on a simplement $Y = X\theta + \varepsilon$ où $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$.

Pour que ce modèle soit identifiable on a besoin de l'hypothèse suivante

Hypothèse 13

On suppose que les colonnes de X_e forment une famille libre. Autrement dit, il n'y a pas de combinaison linéaire reliant les variables explicatives et le vecteur 1_n. De façon équivalente, cela revient à supposer que $\mathrm{rg}(X_e) = p+1$ ou $\mathrm{Ker}(X_e) = \{0_n\}$ ou que la matrice ${}^tX_eX_e$ est inversible.

Remarque: Pour que cette hypothèse soit vérifiée on a forcement besoin que $n \geqslant p+1$. Autrement dit, on a besoin d'au moins plus de données que de paramètres !

C'est une conséquence du fait que, comme $Y \sim \mathcal{N}_n(X_e\theta, \sigma^2 I_n)$, alors si deux jeux de paramètres (θ, σ) et (θ', σ') engendre la même loi on a $X_e\theta = X_e\theta'$ et $\sigma^2 I_n = \sigma'^2 I_n$ d'ou $\sigma = \sigma'$ et

$$X_e\theta = X_e\theta' \;\Rightarrow\; X_e(\theta - \theta') = 0_n \;\Rightarrow\; \theta - \theta' \in \mathrm{Ker}(X_e)$$

et on a donc besoin de supposer que $\mathrm{Ker}(X_e) = \{0_n\}$.

Enfin, comme pour le modèle linéaire simple on définit les résidus du modèle de la façon suivante.

Définition 14

*Soit $\hat{\theta}$ un estimateur de θ. On note $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i,1} + \cdots + \hat{\theta}_p x_{i,p}$ la valeur de y_i estimée par le modèle et $\hat{Y} = X_e\hat{\theta}$ le vecteur des $\hat{y}_i$. On appelle $y_i - \hat{y}_i$ le i-ème **résidu** et $Y - \hat{Y}$ le **vecteur des résidus**.*

2 Estimation des paramètres

a Estimation par maximum de vraisemblance

Proposition 15

Les estimateurs du maximum de vraisemblance de θ et σ^2 existent, sont unique et vérifient

$$\hat{\theta} = ({}^tX_eX_e)^{-1}({}^tX_eY) \;\; \text{et} \;\; \hat{\sigma}^2 = \frac{1}{n}\|Y - \hat{Y}\|_2^2.$$

Afin de calculer le maximum de vraisemblance on va devoir trouver le maximum d'une fonction a plusieurs variables. Pour cela, on va s'aider des résultats suivants.

Lemme 16

- *Soient $v \in \mathbb{R}^d$ et $f : \mathbb{R}^d \to \mathbb{R}$ la fonction définie par $f(x) = \langle v, x\rangle$. Alors, pour tout $x \in \mathbb{R}^d$ on a $\nabla f(x) = v$ et $H(f)(x) = 0_{d\times d}$ où $0_{d\times d}$ est la matrice nulle de taille $d \times d$.*
- *Soient $M \in \mathcal{S}_d(R)$ et $f : \mathbb{R}^d \to \mathbb{R}$ la fonction définie par $f(x) = \langle x, Mx\rangle$. Alors, pour tout $x \in \mathbb{R}^d$ on a $\nabla f(x) = 2Mx$ et $H(f)(x) = 2M$.*

Démonstration :

- On a $f(x) = \sum_{i=1}^d v_i x_i$ donc, pour tout $i, j \in \{1, \cdots, d\}$ on a $\frac{\partial f(x)}{\partial x_i} = v_i$ et $\frac{\partial f(x)}{\partial x_i \partial x_j} = 0$ ce qui prouve le résultat.
- On a $f(x) = \sum_{i,j=1}^d x_i M_{i,j} x_j$. Soit $i \in \{1, \cdots, d\}$, on peut écrire

$$f(x) = x_i^2 M_{i,i} + \sum_{\substack{j=1\\ j\neq i}}^n x_i M_{i,j} x_j + \sum_{\substack{j=1\\ j\neq i}}^n x_j M_{j,i} x_i + \sum_{\substack{j,k=1\\ j,k\neq i}}^n x_j M_{j,k} x_k.$$

En utilisant la symétrie de M on obtient

$$\frac{\partial f(x)}{\partial x_i} = 2x_i M_{i,i} + \sum_{\substack{j=1\\ j\neq i}}^{n} M_{i,j}x_j + \sum_{\substack{j=1\\ j\neq i}}^{n} x_j M_{j,i} = 2\sum_{j=1}^{n} M_{i,j}x_j = 2(Mx)_i$$

$$\text{et} \quad \forall j \in \{1,\cdots,d\}, \ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} = 2M_{i,j}. \qquad \blacksquare$$

Remarque: Ce résultat est l'analogue multidimensionnel du fait que si $f(x) = ax$ alors $f'(x) = a$ et $f''(x) = 0$ et si $f(x) = ax^2$ alors $f'(x) = 2ax$ et $f''(x) = 2a$.

On revient maintenant au calcul des estimateurs du maximum de vraisemblance.

Démonstration de la Proposition 15 : Les observations de Y sont issues d'une loi $\mathcal{N}_n(X_e\theta, \sigma^2 I_n)$ donc, si on note f_{θ,σ^2} la densité de cette loi, on obtient

$$\begin{aligned} L(\theta,\sigma^2|Y) = f_{\theta,\sigma^2}(Y) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det(\sigma^2 I_n)}} \exp\left(-\frac{1}{2}\langle Y - X_e\theta, (\sigma^2 I_n)^{-1}(Y - X_e\theta)\rangle\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|Y - X_e\theta\|^2\right). \end{aligned}$$

La log-vraisemblance du modèle est donc

$$\mathcal{L}(\theta,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - X_e\theta\|^2.$$

Pour un σ fixé, maximiser $\mathcal{L}(\theta,\sigma^2)$ par rapport à θ revient à minimiser la fonction $g(\theta) = \|Y - X_e\theta\|^2$ qui ne dépend pas de σ. L'EMV $\hat{\theta}$ de θ est donc le minimiseur de cette fonction et l'EMV $\hat{\sigma}^2$ de σ^2 est donc le paramètre qui maximise la fonction $\phi : s \mapsto \mathcal{L}(\hat{\theta}, s)$.

Calcul de $\hat{\theta}$:

On cherche θ qui minimise

$$\begin{aligned} g(\theta) = \|Y - X_e\theta\|^2 = \langle Y - X_e\theta, Y - X_e\theta\rangle &= \langle Y, Y\rangle - 2\langle Y, X_e\theta\rangle + \langle X_e\theta, X_e\theta\rangle \\ &= \|Y\|^2 - 2\langle {}^tX_eY, \theta\rangle + \langle \theta, {}^tX_eX_e\theta\rangle. \end{aligned}$$

A l'aide du lemme 16 on peut calculer le gradient et la Hessienne de g par rapport à θ :

$$\nabla g(\theta) = -2{}^tX_eY + 2{}^tX_eX_e\theta \quad \text{et} \quad H(g)(\theta) = 2{}^tX_eX_e.$$

On remarque que $H(g)(\theta)$ est une matrice symétrique semi-définie positive. Par hypothèse on a que ${}^tX_eX_e$ est inversible donc $H(g)(\theta)$ est définie positive et donc g est une fonction strictement convexe. De plus,

$$\nabla g(\theta) = 0 \ \Leftrightarrow \ -2{}^tX_eY + 2{}^tX_eX_e\theta = 0 \ \Leftrightarrow \ \theta = ({}^tX_eX_e)^{-1}({}^tX_eY).$$

La fonction g possède donc un unique minimum global en $\hat{\theta} = ({}^tX_eX_e)^{-1}({}^tX_eY)$.

Calcul de $\hat{\sigma}^2$:

On a

$$\phi'(s) = -\frac{n}{2s} + \frac{1}{2s^2}\|Y - \hat{Y}\|_2^2 = -\frac{n}{2s^2}\left(s - \frac{1}{n}\|Y - \hat{Y}\|_2^2\right).$$

La fonction ϕ possède donc le tableau de variations suivant :

s	0		$\frac{1}{n}\|Y-\hat{Y}\|_2^2$		$+\infty$
$\phi'(s)$		>0	0	<0	
$\phi(s)$		↗		↘	

On en déduit alors que ϕ a un unique maximum en $\hat{\sigma}^2 = \frac{1}{n}\|Y - \hat{Y}\|_2^2$. $\blacksquare$

b Autres méthodes d'estimation

Une façon plus générale de voir le problème de régression linéaire est de retirer l'hypothèse de normalité des erreurs. On considère alors que les données vérifient

$$Y = X_e\theta + \varepsilon \text{ où } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

avec les ε_i i.i.d. centrés et de même variance σ^2 mais pas forcement Gaussien. Dans ce cas-là deux façons classiques d'ajuster le modèle sont les suivantes.

Méthode des moindres carrés

La méthode des moindre carrés consiste à choisir $\hat{\theta}$ afin d'avoir les résidus les plus faibles possibles. On cherche alors à minimiser $\|Y - \hat{Y}\|^2$ et on retrouve

$$\hat{\theta} = ({}^tX_eX_e)^{-1}{}^tX_eY$$

avec les même calculs que pour l'EMV.

Estimateur BLUE

Définition 17

*Un estimateur de θ sous la forme $\hat{\theta} = CY$ pour une certaine matrice C est appelé un **estimateur linéaire**.*

On a $\mathbb{E}[\hat{\theta}] = C\mathbb{E}[Y] = CX_e\theta$ donc $\hat{\theta}$ est non biaisé lorsque $CX_e\theta = \theta$ pour tout θ. Autrement dit, lorsque $CX_e = I_{p+1}$.

Définition 18

*Un estimateur non-biaisé de θ sous la forme $\hat{\theta} = CY$ et dont la variance moyenne est minimal parmi les estimateurs de cette forme est dit **BLUE** (Best linear unbiaised estimator).*

Théorème 19 (*Gauss-Markov*)

L'estimateur $\hat{\theta} = ({}^tX_eX_e)^{-1}({}^tX_eY)$ est BLUE.

Démonstration : Soit $C = ({}^tX_eX_e)^{-1}{}^tX_e + D$ pour une certaine matrice D de taille $(p+1) \times n$. On note $\hat{\theta} = CY$ et $\hat{\theta}' = ({}^tX_eX_e)^{-1}{}^tX_eY$. Comme on impose la condition $CX_e = I_{p+1}$ et que $CX_e = I_{p+1} + DX_e$ alors DX_e est la matrice nulle. On calcule ensuite la matrice de variance de $\hat{\theta}$:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(CY) \\
&= C\mathrm{Var}(Y){}^tC \\
&= \sigma^2 C{}^tC \\
&= \sigma^2(({}^tX_eX_e)^{-1}{}^tX_e + D)(X_e({}^tX_eX_e)^{-1} + {}^tD) \\
&= \sigma^2(({}^tX_eX_e)^{-1}{}^tX_eX_e({}^tX_eX_e)^{-1} + DX_e({}^tX_eX_e)^{-1} + ({}^tX_eX_e)^{-1}{}^tX_e{}^tD + D{}^tD) \\
&= \sigma^2(({}^tX_eX_e)^{-1} + D{}^tD)
\end{aligned}$$

Comme l'estimateur $\hat{\theta}'$ correspond au cas où D est la matrie nulle alors $\mathrm{Var}(\hat{\theta}') = \sigma^2({}^tX_eX_e)^{-1}$ et donc

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}') + \sigma^2 D{}^tD.$$

Le variance moyenne des $\hat{\theta}_i$ est donc

$$\frac{1}{p+1}\sum_{i=0}^{p}\mathrm{Var}(\hat{\theta}_i) = \frac{1}{p+1}\mathrm{Tr}(\mathrm{Var}(\hat{\theta})) = \frac{1}{p+1}\mathrm{Tr}(\mathrm{Var}(\hat{\theta}')) + \frac{\sigma^2}{p+1}\mathrm{Tr}(D^tD)$$
$$= \frac{1}{p+1}\sum_{i=0}^{p}\mathrm{Var}(\hat{\theta}'_i) + \frac{\sigma^2}{p+1}\mathrm{Tr}(D^tD)$$

et comme D^tD est semi-définie positive alors $\mathrm{Tr}(D^tD) \geqslant 0$ ce qui prouve le résultat. ■

Remarques:

- Dû à ce résultat, l'estimateur $\hat{\theta} = ({}^tX_eX_e)^{-1}({}^tX_eY)$ est souvent appelé l'estimateur de Gauss-Markov.
- Ces résultats montrent que l'estimateur de θ obtenu par maximum de vraisemblance reste un bon estimateur même sans l'hypothèse de normalité des erreurs.

3 Loi des estimateurs

Proposition 20

$$\hat{\theta} \sim \mathcal{N}_{p+1}(\theta, \sigma^2({}^tX_eX_e)^{-1})$$

Démonstration : Comme $Y = X_e\theta + \varepsilon$ alors

$$\hat{\theta} = ({}^tX_eX_e)^{-1}({}^tX_eY) = ({}^tX_eX_e)^{-1}({}^tX_e(X_e\theta+\varepsilon)) = \theta + ({}^tX_eX_e)^{-1}({}^tX_e\varepsilon).$$

Or, $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ donc $({}^tX_eX_e)^{-1}({}^tX_e\varepsilon) \sim \mathcal{N}_n(0_n, \Sigma)$ où

$$\Sigma = ({}^tX_eX_e)^{-1}({}^tX_e)(\sigma^2 I_n)^t(({}^tX_eX_e)^{-1}({}^tX_e))$$
$$= \sigma^2({}^tX_eX_e)^{-1}({}^tX_eX_e)({}^tX_eX_e)^{-1} = \sigma^2({}^tX_eX_e)^{-1}.$$

On obtient donc bien $\hat{\theta} \sim \mathcal{N}_{p+1}(\theta, \sigma^2({}^tX_eX_e)^{-1})$. ■

Remarques:

- $\hat{\theta}$ est non biaisé.
- Si on se place dans le cas où on a $p = 1$ variable explicative quantitative $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ alors

$X_e = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ et donc

$${}^tX_eX_e = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$
$$\Rightarrow ({}^tX_eX_e)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$
$$\Rightarrow ({}^tX_eX_e)^{-1} = \frac{1}{n\mathrm{var}(X)}\begin{pmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}$$

et on retrouve les résultats du modèle linéaire simple.

- Si $p = 1$ alors ${}^tX_eX_e$ est inversible ssi $\text{var}(X) > 0$ ce qui est bien l'hypothèse du modèle linéaire simple.

Afin d'établir la loi de σ^2 on établit d'abord le résultat général suivant que l'on utilisera plusieurs fois par la suite. On rappelle qu'on note P_E la projection orthogonal sur un ensemble E et que la projection orthogonale sur l'espace engendré par les colonnes d'une matrice M s'écrit $M({}^tMM)^{-1}{}^tM$.

Proposition 21

Soit E l'espace vectoriel engendré par les colonnes de X_e et qui contient donc Vect(1_n). On pose G le complément orthogonal de Vect(1_n) dans E. Alors :

$$Y - \hat{Y} = P_{E^\perp}(Y),\ \hat{Y} - \bar{Y}1_n = P_G(Y) \text{ et } Y - \bar{Y}1_n = P_{Vect_{(1_n)^\perp}}(Y).$$

Démonstration : On a

$$\hat{Y} = X_e\hat{\theta} = X_e({}^tX_eX_e)^{-1}{}^tX_eY = P_E(Y)$$

d'où $Y - \hat{Y} = P_{E^\perp}(Y)$. De plus, on a

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\langle Y, 1_n\rangle}{\langle 1_n, 1_n\rangle} \ \Rightarrow\ \bar{Y}1_n = P_{\text{Vect}_{(1_n)}}(Y)$$

d'où $\hat{Y} - \bar{Y}1_n = P_G(Y)$ et $Y - \bar{Y}1_n = P_{\text{Vect}_{(1_n)^\perp}}(Y)$. ■

Corollaire 22

$$\|Y - \bar{Y}1_n\|_2^2 = \|Y - \hat{Y}\|_2^2 + \|\hat{Y} - \bar{Y}1_n\|_2^2$$

Démonstration : Comme $E = \text{Vect}(1_n) \overset{\perp}{\oplus} G$ alors $\mathbb{R}^n = E \overset{\perp}{\oplus} E^\perp = \text{Vect}(1_n) \overset{\perp}{\oplus} G \overset{\perp}{\oplus} E^\perp$ donc $\text{Vect}(1_n)^\perp = G \overset{\perp}{\oplus} E^\perp$ et par le théorème de Pythagore on en déduit

$$\|P_{\text{Vect}_{(1_n)^\perp}}(Y)\|_2^2 = \|P_G(Y)\|_2^2 + \|P_{E^\perp}(Y)\|_2^2,$$

d'où le résultat. ■

Remarque:

$$\|Y - \bar{Y}1_n\|_2^2 = \sum_{i=1}^{n}(y_i - \bar{Y})^2 = n\text{var}(Y).$$

$\|Y - \bar{Y}1_n\|_2^2$ correspond aussi à l'inertie totale des y_i. $\|Y - \hat{Y}\|_2^2$ est appelé **l'inertie des résidus** et $\|\hat{Y} - \bar{Y}1_n\|_2^2$ est appelé **l'inertie expliquée par la régression linéaire**. Cela motive la définition suivante d'un critère de qualité du modèle.

Définition 23

On définit le R^2 d'un modèle linéaire par la proportion de l'inertie totale expliquée par la régression linéaire :

$$R^2 = \frac{\|\hat{Y} - \bar{Y}1_n\|_2^2}{\|Y - \bar{Y}1_n\|_2^2} = 1 - \frac{\|Y - \hat{Y}\|_2^2}{n\text{var}(Y)} \in [0, 1].$$

Remarque: Plus le R^2 est proche de 1, plus le modèle est bien ajusté.

Proposition 24

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1)).$$

De plus, $\hat{\theta}$ est indépendant de $\hat{\sigma}^2$.

Démonstration : On a $Y = X_e\theta + \varepsilon$ et $X_e\theta \in \mathrm{Im}(X_e) = E$ donc

$$Y - \hat{Y} = P_{E^\perp}(Y) = P_{E^\perp}(\varepsilon).$$

Or, par hypothèse la dimension de E est $p+1$ donc la dimension de $E^\perp$ est $n-(p+1)$ donc par le théorème de Cochran on en déduit

$$\frac{\|Y-\hat{Y}\|_2^2}{\sigma^2} = \frac{\|P_{E^\perp}(\varepsilon)\|_2^2}{\sigma^2} \sim \chi^2(n-(p+1))$$

et on en déduit que $Y-\hat{Y}$, et donc $\hat{\sigma}^2$, est indépendant de $P_E(\varepsilon) = \hat{Y} = X_e\hat{\theta}$ et donc de $\hat{\theta}$. ■

Remarque: On a $\mathbb{E}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = n-(p+1)$ donc $\mathbb{E}[\hat{\sigma}^2] = \frac{n-(p+1)}{n}\sigma^2$ et donc $\hat{\sigma}^2$ est biaisé.

On définit alors une version corrigée de l'estimateur de σ^2.

Définition 25

On définit

$$\widetilde{\sigma}^2 = \frac{n}{n-(p+1)}\hat{\sigma}^2 = \frac{1}{n-(p+1)}\|Y-\hat{Y}\|_2^2$$

qui est un estimateur non biaisé de σ^2. Il vérifie

$$\frac{(n-(p+1))\widetilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$$

et est indépendant de $\hat{\theta}$.

4 Intervalles de confiance et tests pour le modèle

a Intervalles de confiance et tests de nullité des $\hat{\theta}_i$

On a déjà vu que $\hat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2\Sigma_{i+1,i+1})$ avec $\Sigma = ({}^tX_eX_e)^{-1}$ pour $i \in \{0,\cdots,p\}$ et que $\widetilde{\sigma}^2$ est indépendant de $\hat{\theta}_i$. On a donc

$$\frac{\hat{\theta}_i-\theta_i}{\sqrt{\sigma^2\Sigma_{i+1,i+1}}} \sim \mathcal{N}(0,1) \quad \text{et} \quad \frac{(n-(p+1))\widetilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$$

donc

$$\frac{\frac{\hat{\theta}_i-\theta_i}{\sqrt{\sigma^2\Sigma_{i+1,i+1}}}}{\sqrt{\frac{(n-(p+1))\widetilde{\sigma}^2}{(n-(p+1))\sigma^2}}} = \frac{\hat{\theta}_i-\theta_i}{\sqrt{\widetilde{\sigma}^2\Sigma_{i+1,i+1}}} \sim \mathcal{T}(n-(p+1)).$$

On obtient donc un intervalle de confiance exact de risque α pour θ_i :

$$IC_{1-\alpha}(\theta_i) = \left[\hat{\theta}_i \pm t_{1-\alpha/2}^{(n-(p+1))}\sqrt{\widetilde{\sigma}^2\Sigma_{i+1,i+1}}\right].$$

De plus, si on souhaite tester l'hypothèse $\mathcal{H}_0 : \theta_i = 0$ contre $\mathcal{H}_1 : \theta_i \neq 0$ alors on pose la statistique

$$T = \frac{\hat{\theta}_i}{\sqrt{\widetilde{\sigma}^2\Sigma_{i+1,i+1}}}.$$

Sous $\mathcal{H}_0$ on a en général $T \sim \mathcal{T}(n-(p+1))$ et sous $\mathcal{H}_1$ on a $|T| \to +\infty$. On rejette donc $\mathcal{H}_0$ lorsque $|T| \geqslant t_{1-\alpha/2}^{(n-(p+1))}$.

b Test de nullité de combinaisons linéaires de paramètres

Maintenant, on suppose que l'on cherche à tester la nullité d'une (ou plusieurs) combinaisons linéaires non liées de paramètres. Les hypothèses du test s'écrivent alors sous la forme

$$\mathcal{H}_0 : C\theta = 0_k \quad \text{contre} \quad \mathcal{H}_1 : C\theta \neq 0_k,$$

où C est une matrice de taille $k \times (p+1)$ de rang plein.

Exemple: Si $p = 2$ et qu'on souhaite tester si $\theta_0 + \theta_1 = 0$ et $\theta_2 - \theta_1 = 0$ en même temps alors cela revient à tester

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Pour faire ce test on va avoir besoin de la loi de Fisher.

Définition 26

Soit $X \sim \chi^2(n)$ et $Y \sim \chi^2(m)$ deux variables aléatoires indépendantes. On appelle ***loi de Fisher*** *à n degrés de liberté au numérateur et m degrés de liberté au dénominateur, notée $\mathcal{F}(n,m)$, la loi sur $[0,+\infty[$ de la variable*

$$\frac{X/n}{Y/m}.$$

Proposition 27

Sous $\mathcal{H}_0$ la statistique

$$F = \frac{\langle C\hat{\theta}, (C\Sigma^t C)^{-1} C\hat{\theta}\rangle}{k\widetilde{\sigma}^2}$$

suit une loi de Fisher $\mathcal{F}(k, n-(p+1))$.

Démonstration : On sait déjà que $\widetilde{\sigma}^2$ et $\hat{\theta}$ sont indépendants et que

$$\frac{(n-(p+1))\widetilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1)).$$

De plus,

$$\langle C\hat{\theta}, (C\Sigma^t C)^{-1} C\hat{\theta}\rangle = \langle (C\Sigma^t C)^{-1/2} C\hat{\theta}, (C\Sigma^t C)^{-1/2} C\hat{\theta}\rangle = \|(C\Sigma^t C)^{-1/2} C\hat{\theta}\|_2^2,$$

ce qui explique pourquoi F est à valeur positive. Or, on a $\hat{\theta} \sim \mathcal{N}_{p+1}(\theta, \sigma^2\Sigma)$ donc

$$\begin{aligned}(C\Sigma^t C)^{-1/2} C\hat{\theta} \sim & \mathcal{N}_k((C\Sigma^t C)^{-1/2} C\theta, \sigma^2 (C\Sigma^t C)^{-1/2} C\Sigma^t ((C\Sigma^t C)^{-1/2} C)) \\ \sim & \mathcal{N}_k((C\Sigma^t C)^{-1/2} C\theta, \sigma^2 (C\Sigma^t C)^{-1/2} C\Sigma^t C (C\Sigma^t C)^{-1/2}) \\ \sim & \mathcal{N}_k((C\Sigma^t C)^{-1/2} C\theta, \sigma^2 I_k).\end{aligned}$$

Sous $\mathcal{H}_0$ on a $C\theta = 0_k$ donc $(C\Sigma^t C)^{-1/2} C\hat{\theta} \sim \mathcal{N}_k(0, \sigma^2 I_k)$ d'où

$$\frac{\|(C\Sigma^t C)^{-1/2} C\hat{\theta}\|_2^2}{\sigma^2} \sim \chi^2(k).$$

En conséquence, on en déduit que

$$\frac{\frac{\|(C\Sigma^t C)^{-1/2}C\hat{\theta}\|_2^2}{k\sigma^2}}{\frac{(n-(p+1))\widetilde{\sigma}^2}{(n-(p+1))\sigma^2}} \sim \mathcal{F}(k, n-(p+1))$$

et en simplifiant la fraction ça nous donne le résultat attendu. ■

On remarque que sous $\mathcal{H}_1$ la variable $(C\Sigma^t C)^{-1/2}C\hat{\theta}$ a pour espérance $(C\Sigma^t C)^{-1/2}C\theta$ qui est non nul. En général F a tendance à prendre de grande valeurs sous $\mathcal{H}_1$. On rejette donc $\mathcal{H}_0$ au risque α lorsque F est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(k, n-(p+1))$. C'est ce qu'on appelle le **test de Wald**.
Exemple: Si on reprend les 433 données de taille de femme de Sir Francis Galton alors on peut considérer deux modèles de dépendance entre la taille des enfants Y_i, celle de la mère x_i et celle du père z_i :

- Un modèle linéaire multiple entre toutes les variables :

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 z_i + \varepsilon_i.$$

- Un modèle linéaire simple entre la taille des enfants et la taille moyenne des parents :

$$Y_i = \mu_0 + \mu_1(x_i + z_i)/2 + \varepsilon_i.$$

On remarque que si $\theta_1 = \theta_2$ alors le premier modèle est équivalent au deuxième en prenant $\mu_0 = \theta_0$ et $\mu_1 = 2\theta_1 = 2\theta_2$. On veut donc tester l'hypothèse $\theta_1 = \theta_2$ pour ce premier modèle ce qui revient à faire un test de Wald avec $C = (0, 1, -1)$. On trouve alors $F \approx 2.7$. Or, le quantile d'ordre 0.95 de la loi $\mathcal{F}(1, 433-(2+1))$ est ≈ 3.86 donc on ne rejette pas l'hypothèse que $\theta_1 = \theta_2$ et donc qu'on peut utiliser le deuxième modèle plus simple.

c Intervalle de prédiction

On considère que l'on observe une nouvelle valeur $(x_{n+1,1}, \cdots, x_{n+1,p})$ des variables explicatives mais que l'on ne connaît pas la valeur y_{n+1} de la variable d'intérêt associée. On considère que c'est la réalisation d'une variable aléatoire

$$Y_{n+1} = \theta_0 + \sum_{j=1}^{p} \theta_j x_{n+1,j} + \varepsilon_{n+1}$$

où ε_{n+1} est indépendant et de même loi que les ε_i. On va donc estimer y_{n+1} par

$$\hat{y}_{n+1} = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j x_{n+1,j}.$$

Proposition 28

*Soit $v = (1, x_{n+1,1}, \cdots, x_{n+1,p})$. Alors un intervalle de confiance exact de risque α pour Y_{n+1}, appelé **intervalle de prédiction**, est*

$$\left[\hat{y}_{n+1} \pm t_{1-\alpha/2}^{(n-(p+1))}\widetilde{\sigma}\sqrt{1 + v({}^tX_eX_e)^{-1}{}^tv}\right].$$

Démonstration : On a $\hat{y}_{n+1} = v\hat{\theta}$ et donc $\hat{y}_{n+1} - Y_{n+1} = v(\hat{\theta} - \theta) + \varepsilon_{n+1}$. Comme $\hat{\theta} \sim \mathcal{N}_{p+1}(\theta, \sigma^2({}^tX_eX_e)^{-1})$ alors

$$v(\hat{\theta} - \theta) \sim \mathcal{N}(0, \sigma^2 v({}^tX_eX_e)^{-1t}v).$$

Comme $v(\hat{\theta} - \theta)$ dépend juste de $\hat{\theta}$ et donc de ε alors $v(\hat{\theta} - \theta)$ est indépendant de ε_{n+1} d'où

$$\hat{y}_{n+1} - Y_{n+1} = v(\hat{\theta} - \theta) + \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2(1 + v({}^tX_eX_e)^{-1t}v))$$

et donc

$$\frac{\hat{y}_{n+1} - Y_{n+1}}{\sigma\sqrt{1 + v({}^tX_eX_e)^{-1t}v}} \sim \mathcal{N}(0, 1).$$

On a vu que $\widetilde{\sigma}$ est indépendant de $\hat{\theta}$. De plus, comme $\widetilde{\sigma}$ est construit à partir de ε alors il est aussi indépendant de ε_{n+1} donc $\widetilde{\sigma}$ est indépendant de $\hat{y}_{n+1} - Y_{n+1}$. On en déduit alors :

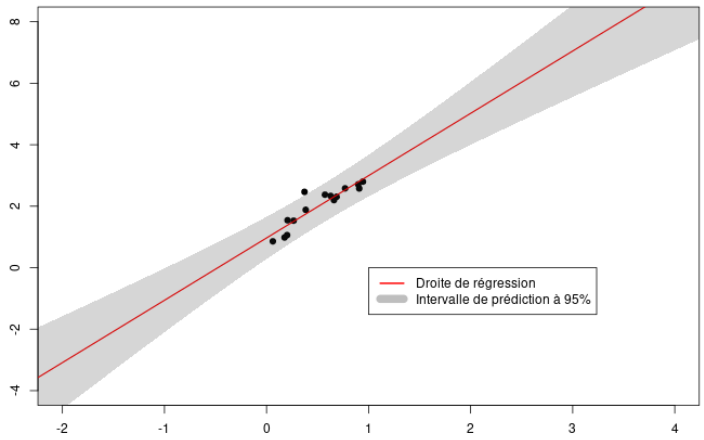
$$\frac{\frac{\hat{y}_{n+1} - Y_{n+1}}{\sigma\sqrt{1+v({}^tX_eX_e)^{-1t}v}}}{\sqrt{\frac{(n(p+1))\widetilde{\sigma}^2}{(n(p+1))\sigma^2}}} \sim T(n - (p+1)) \;\Rightarrow\; \frac{\hat{y}_{n+1} - Y_{n+1}}{\widetilde{\sigma}\sqrt{1 + v({}^tX_eX_e)^{-1t}v}} \sim \mathcal{T}(n - (p+1))$$

et donc

$$\mathbb{P}\left(\left|\frac{\hat{y}_{n+1} - Y_{n+1}}{\widetilde{\sigma}\sqrt{1 + v({}^tX_eX_e)^{-1t}v}}\right| \leqslant t_{1-\alpha/2}^{(n-(p+1))}\right) = 1 - \alpha$$

d'où en découle l'intervalle de prédiction. ■

Exemple: On donne çi-dessous une illustration de cet intervalle de prédiction pour la régression linéaire simple sur un petit jeu de données. On peut voir que la prédiction est de plus en plus imprécise au fur à mesure que l'on essaye de prédire ce qui se passe sur des données très différentes de celles du jeu de données.



5 Qualité d'ajustement du modèle

a Comparaison au modèle vide

Un cas particulier important des tests de Wald est le test de l'hypothèse $\theta_1 = \cdots = \theta_p = 0$. Autrement dit, on teste s'il y a une différence significative entre le modèle qui prend en compte

toute les variables explicatives et le modèle qui les ignore toutes et possède juste un intercept (appelé le **modèle vide**). Cela revient donc à tester la nullité de $C\theta$ avec

$$C = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

qui est de taille $p \times (p+1)$. La statistique F correspondante sera donc comparée aux quantiles de la loi $\mathcal{F}(p, n-(p+1))$.

b Le R^2 ajusté

On a vu qu'un critère de qualité du modèle qui apparaît naturellement est le R^2 :

$$R^2 = 1 - \frac{\|Y - \hat{Y}\|^2}{n\text{var}(Y)} \in [0,1].$$

A noter que dans le cas de la régression linéaire simple alors le R^2 correspond à la corrélation entre X et Y :

Proposition 29

Si on se place dans le cas où on a $p = 1$ variable explicative quantitative $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ alors

$$R^2 = corr(X,Y)^2.$$

Démonstration : En reprenant les notations du modèle linéaire simple on a $\hat{Y} = \hat{\alpha}X + \hat{\beta}1_n$. Or, $\hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X}$ d'où

$$\begin{aligned} \|Y - \hat{Y}\|^2 &= \|(Y - \bar{Y}1_n) - \hat{\alpha}(X - \bar{X}1_n)\|^2 \\ &= \|Y - \bar{Y}1_n\|^2 + \hat{\alpha}^2\|X - \bar{X}1_n\|^2 - 2\hat{\alpha}\langle Y - \bar{Y}1_n, X - \bar{X}1_n\rangle \\ &= \text{var}(Y) + \frac{\text{cov}(X,Y)^2}{\text{var}(X)^2}\text{var}(X) - 2\frac{\text{cov}(X,Y)}{\text{var}(X)}\text{cov}(X,Y) \\ &= \text{var}(Y) - \frac{\text{cov}(X,Y)^2}{\text{var}(X)} \end{aligned}$$

et donc

$$1 - \frac{\|Y - \hat{Y}\|^2}{n\text{var}(Y)} = \frac{\text{cov}(X,Y)^2}{\text{var}(X)\text{var}(Y)} = \text{corr}(X,Y)^2.$$

■

Plus R^2 est proche de 1 plus le modèle est bien ajusté. Néanmoins, le R^2 a tendance a naturellement augmenter avec le nombre de variables explicatives. C'est dû au fait que

$$\mathbb{E}[n\text{var}(Y)] = (n-1)\sigma^2 \text{ et } \mathbb{E}[\|Y - \hat{Y}\|_2^2] = (n-(p+1))\sigma^2.$$

Du coup, quand p augmente alors $\|Y - \hat{Y}\|_2^2$ va diminuer mais cela ne vas pas affecter $\text{var}(Y)$ ce qui va entraîner une augmentation du R^2. Afin de compenser ce phénomène on utilise une version modifiée du R^2 appelée le R^2 **ajusté**.

Définition 30

*On définit le R^2 **ajusté** d'un modèle linéaire par*

$$R^2_{\text{ajusté}} = 1 - \frac{\|Y - \hat{Y}\|^2/(n-(p+1))}{n\text{var}(Y)/(n-1)}$$

Remarque: Le R^2 ajusté est tout le temps plus petit que le R^2 et il peut prendre des valeurs négatives.

c Analyse des résidus

Lorsqu'on fait une régression linéaire on considère plusieurs hypothèses :

- $\mathbb{E}[Y_i]$ est une fonction linéaire des variables explicatives.
- Les termes d'erreur ε_i sont additifs.
- Les termes d'erreur ε_i sont indépendants.
- Les termes d'erreur ε_i sont de loi normale.
- Les termes d'erreur ε_i ont la même variance.

Il existe plusieurs moyens de voir si ces hypothèses ont l'air d'être vérifiée sur les données en s'intéressant au comportement des résidus $y_i - \hat{y}_i$ comme estimateur des ε_i. Par exemple, comme on a vu que $Y - \hat{Y}$ est indépendant de $\hat{\theta}$, et donc de $\hat{Y} = X_e\hat{\theta}$, alors on peut regarder le nuage de points des résidus $y_i - \hat{y}_i$ en fonction des valeurs prédites $\hat{y}_i$. Normalement, il ne devrait pas y avoir de comportement particulier apparaître. Un autre outil utile est les résidus renormalisés par un estimateur de leur écart-type appelé les **résidus standardisés**

Proposition 31

Le vecteur des résidus $Y - \hat{Y}$ vérifie

$$Y - \hat{Y} \sim \mathcal{N}_n(0_n, \sigma^2(I_n - X_e({}^tX_eX_e)^{-1}{}^tX_e)).$$

Démonstration : On a vu dans la preuve de la proposition 21 que

$$Y - \hat{Y} = (I_n - X_e({}^tX_eX_e)^{-1}{}^tX_e)\varepsilon.$$

Si on pose $P = (I_n - X_e({}^tX_eX_e)^{-1}{}^tX_e)$ qui est une matrice symétrique de projection alors

$$Y - \hat{Y} \sim \mathcal{N}_n(P0_n, P(\sigma^2 I_n)P^T) \;\Rightarrow\; Y - \hat{Y} \sim \mathcal{N}_n(0_n, \sigma^2 P).$$

■

Il en vient deux façons naturelles de renormaliser les résidus.

Définition 32

*Soit h_i la i-ème valeur de la diagonale de la matrice $X_e({}^tX_eX_e)^{-1}{}^tX_e$ appelé **l'effet de levier** (**leverage score** en anglais). On définit le i-ème **résidu standardisé** (ou **studentisé interne**) par*

$$\frac{Y_i - \hat{Y}_i}{\tilde{\sigma}\sqrt{1-h_i}}.$$

*On définit le i-ème résidu **studentisé externe** par*

$$\frac{Y_i - \hat{Y}_i}{\tilde{\sigma}_{-i}\sqrt{1-h_i}} \sim \mathcal{T}(n-(p+1)),$$

où $\tilde{\sigma}_{-i}$ est l'estimation de sigma obtenue par la régression linéaire sur tous les individus sauf le i-ème.

Remarque: Comme $\tilde{\sigma}$ et $Y_i - \hat{Y}_i$ dépendent de ε_i alors ces deux quantités ne sont pas forcement indépendantes. C'est pour ça qu'on ne connaît pas la loi des résidus standardisés. Néanmoins, si n est assez grand, cette loi est proche d'une loi de Student (et donc d'une loi normale). A l'inverse, on connaît la loi exacte des résidus studentisés externe mais pour les utiliser on doit calculer tous les $\tilde{\sigma}_{-i}$ ce qui peut prendre du temps.

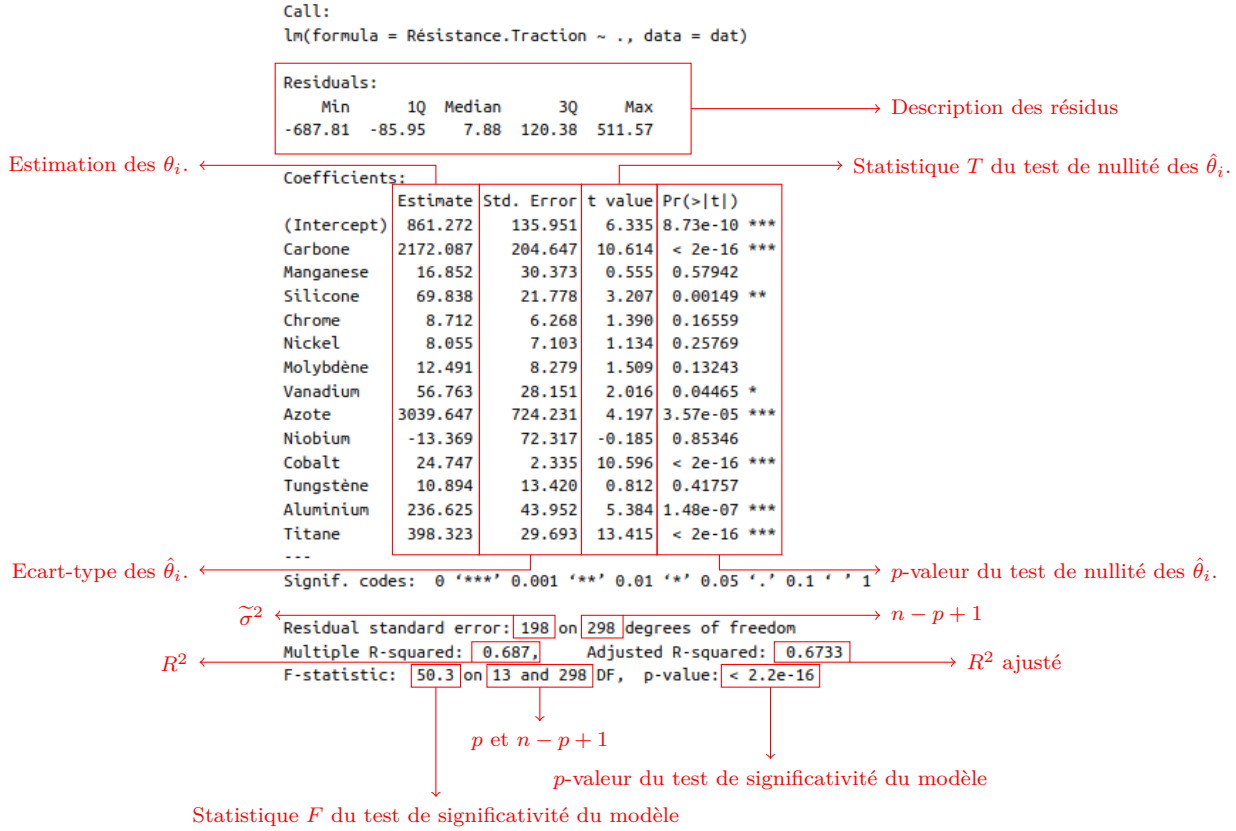
On peut donc regarder un QQ-plot de ces résidus standardisés et si le modèle est bon on doit voir quelque chose similaire à une loi normale. On peut aussi regarder un plot des résidus standardisés en fonction des valeurs prédites et il ne doit pas apparaître comportements particuliers.

6 Exemple d'application sur R

Comme exemple on utilisera des données de composition et de propriétés mécaniques de $n = 312$ aciers. On s'intéresse à modéliser la résistance à la traction (en mégapascal) de ces aciers, c'est à dire la force d'étirement nécessaire pour les casser, en fonction de leurs compositions (en %) de $p = 13$ atomes : C, Mn, Si, Cr, Ni, Mo, V, N, Nb, Co, W, Al, Ti.

C	Mn	Si	Cr	Ni	Mo	V	N	Nb	Co	W	Al	Ti	Résist. Traction
0.02	0.05	0.05	0.01	19.70	2.95	0.01	0.00	0.01	15.00	0.00	0.15	1.55	2473.50
0.18	0.01	0.01	13.44	0.01	3.01	0.46	0.04	0.01	19.46	2.35	0.04	0.00	1929.20
0.00	0.01	0.01	8.67	13.45	0.82	0.01	0.00	0.01	13.90	0.00	0.39	0.57	1871.80
0.01	0.05	0.05	0.01	17.70	3.95	0.01	0.00	0.01	15.00	0.00	0.13	1.47	2514.90
0.01	0.05	0.05	0.01	19.40	1.45	0.01	0.00	0.01	14.90	0.00	0.13	1.55	2315.00
0.19	0.02	0.49	12.56	0.94	1.96	0.01	0.00	0.01	20.10	0.00	0.03	0.00	1779.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

On montre çi-dessous le résultat d'une régression linéaire sur R de la résistance à la traction en fonction des proportions de chaque atome dans l'acier.



On peut faire les remarques suivantes :

- Le modèle est assez bien ajusté (R^2 ajusté de 0.67) et il est significativement meilleur que le modèle vide (p-valeur $< 2.2 \times 10^{-16}$).
- Les variables ayant un effet significatif dans le modèle sur la résistance à la traction de l'acier sont : la proportion en carbone (p-valeur $< 2.2 \times 10^{-16}$), la proportion en silicone

(p-valeur $= 0.0015$), la proportion en vanadium (p-valeur $= 0.045$), la proportion en azote (p-valeur $= 3.6 \times 10^{-5}$), la proportion en cobalt (p-valeur $< 2.2 \times 10^{-16}$), la proportion en aluminium (p-valeur $= 1.48 \times 10^{-7}$) et la proportion en titane (p-valeur $< 2.2 \times 10^{-16}$). L'intercept est aussi significatif (p-valeur $= 8.7 \times 10^{-10}$).

- Toutes ces variables ont un coefficient positif. Ca veut dire que, a proportion fixée pour les autres éléments, une augmentation de la proportion de ces atomes augmente la résistance à la traction de l'acier.

IV Analyse de la variance (ANOVA)

1 Écriture du modèle

On considère toujours que la variable d'intérêt Y est quantitative mais on suppose maintenant qu'on a une unique variable explicative X qui est qualitative avec $J \geqslant 2$ modalités qu'on numérote 1 à J. Le modèle ANOVA consiste à supposer que lorsque $x_i = j$ alors y_i est la réalisation d'une variable aléatoire $Y_i \sim \mathcal{N}(\theta_j, \sigma^2)$ avec indépendance entre les Y_i.
Si on note $\mathbb{1}_{i,j}$ la variable qui vaut 1 si l'individu i possède la modalité j et 0 sinon alors on a

$$Y_i = \theta_1 \mathbb{1}_{i,1} + \cdots + \theta_I \mathbb{1}_{i,J} + \varepsilon_i \quad \text{où} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

On retrouve bien un modèle linéaire pour les variables $\mathbb{1}_{\bullet,j} = \begin{pmatrix} \mathbb{1}_{1,j} \\ \vdots \\ \mathbb{1}_{n,j} \end{pmatrix}$ mais sans intercept.

Remarque: Comme $\mathbb{1}_{i,1} + \cdots + \mathbb{1}_{i,J} = 1$ alors il est aussi possible de faire apparaître un intercept en posant

$$\begin{cases} \mu_0 = \theta_J \\ \mu_1 = \theta_1 - \theta_J \\ \vdots \\ \mu_{J-1} = \theta_{J-1} - \theta_J \end{cases} \Leftrightarrow \begin{cases} \theta_1 = \mu_1 + \mu_0 \\ \vdots \\ \theta_{J-1} = \mu_{J-1} + \mu_0 \\ \theta_J = \mu_0 \end{cases}$$

ce qui donne

$$\begin{aligned} Y_i &= \theta_1 \mathbb{1}_{i,1} + \cdots + \theta_J \mathbb{1}_{i,J} + \varepsilon_j \\ &= (\mu_1 + \mu_0)\mathbb{1}_{i,1} + \cdots + (\mu_{J-1} + \mu_0)\mathbb{1}_{i,J-1} + \mu_0 \mathbb{1}_{i,J} + \varepsilon_j \\ &= \mu_0 + \mu_1 \mathbb{1}_{i,1} + \cdots + \mu_{J-1}\mathbb{1}_{i,J-1}\varepsilon_j \end{aligned}$$

Dans ce cas-là, une modalité est mise de côté. Elle est appelée **la modalité de référence**. Les coefficients associés aux variables correspondent alors à la différence entre les vrais coefficients et celui de la modalité de référence.

La façon classique de traiter le modèle ANOVA est de regrouper les individus selon les modalités qu'ils possèdent. On note n_j le nombre d'individus possédant la modalité j et $y_{i,j}$ la valeur de la variable Y pour le i-ème individu possédant la modalité j. $y_{i,j}$ est donc la simulation d'une variable $Y_{i,j} \sim \mathcal{N}(\theta_j, \sigma^2)$. Le modèle ANOVA s'écrit alors sous la forme

$$Y = X\theta + \varepsilon.$$

avec

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{n_1,1} \\ Y_{1,2} \\ \vdots \\ Y_{n_2,2} \\ \vdots \\ Y_{1,J} \\ \vdots \\ Y_{n_J,J} \end{pmatrix}, \; X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \; \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_J \end{pmatrix} \text{ et } \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n).$$

Définition 33

On note $\hat{y}_{i,j} = \hat{\theta}_i$ la valeur de $y_{i,j}$ estimée par le modèle. On note alors $Y - \hat{Y}$ le vecteur des résidus avec

$$\hat{Y} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_J \\ \vdots \\ \hat{\theta}_J \end{pmatrix} = X\hat{\theta}$$

Définition 34

On note $\overline{Y_{\bullet,j}} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{i,j}$ la moyenne des $Y_{i,j}$ pour $i \in \{1, \cdots, n_j\}$ (la moyenne des valeurs de Y pour les individus avec la modalité j) et $\overline{Y}$ la moyenne de tous les $Y_{i,j}$.

2 Estimation et loi des paramètres

Proposition 35

On suppose que $n_j \geqslant 1$ pour tout j.

- *L'estimateur du maximum de vraisemblance de chaque θ_j est $\hat{\theta}_j = \overline{Y_{\bullet,j}}$. De plus, les $\hat{\theta}_j$ sont indépendants de loi $\hat{\theta}_j \sim \mathcal{N}\left(\theta_j, \frac{\sigma^2}{n_j}\right)$*
- *L'estimateur du maximum de vraisemblance de σ^2 est $\hat{\sigma}^2 = \frac{1}{n}\|Y - \hat{Y}\|_2^2$. De plus, $\hat{\sigma}^2$ est indépendant de $\hat{\theta}$ et*

$$n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - J).$$

Démonstration : Le résultat sur $\hat{\sigma}$ est une conséquence directe du résultat sur les modèles li-

néaires multiples. Pour les $\hat{\theta}_i$ on cherche à calculer $\hat{\theta} = ({}^tXX)^{-1}{}^tXY$. On a

$$ {}^tXX = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} = \begin{pmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_I \end{pmatrix} $$

qui est bien inversible d'inverse $({}^tXX)^{-1} = \begin{pmatrix} 1/n_1 & & 0 \\ & \ddots & \\ 0 & & 1/n_I \end{pmatrix}$. De plus,

$$ {}^tXY = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ \vdots \\ y_{n_2,2} \\ \vdots \\ y_{1,J} \\ \vdots \\ y_{n_J,J} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n_1} y_{i,1} \\ \sum_{i=1}^{n_2} y_{i,2} \\ \vdots \\ \sum_{i=1}^{n_J} y_{i,J} \end{pmatrix} $$

et donc

$$ \hat{\theta} = ({}^tXX)^{-1}{}^tXY = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i,1} \\ \vdots \\ \frac{1}{n_J}\sum_{i=1}^{n_J} y_{i,J} \end{pmatrix} = \begin{pmatrix} \overline{Y_{\bullet,1}} \\ \vdots \\ \overline{Y_{\bullet,J}} \end{pmatrix} \Rightarrow \forall j \in \{1, \cdots, J\},\ \hat{\theta}_j = \overline{Y_{\bullet,j}}. $$

De plus, comme $\hat{\theta} \sim \mathcal{N}_J(\theta, \sigma^2({}^tXX)^{-1})$ et $({}^tXX)^{-1}$ est diagonale alors les $\hat{\theta}_j$ sont indépendants et de variance σ^2/n_j. ■

3 Le test ANOVA

Le modèle ANOVA est en général utilisé lorsqu'on cherche s'il y a une différence significative entre la moyenne des individus séparés par modalités. Autrement dit, l'hypothèse que l'on souhaite tester est

$$ \mathcal{H}_0 : \{\theta_1 = \cdots = \theta_J\} \text{ contre } \mathcal{H}_1 : \{\exists i, j \text{ t.q. } \theta_i \neq \theta_j\}. $$

Remarque: On peut réécrire l'hypothèse $\mathcal{H}_0$ comme la nullité de $J-1$ combinaison linéaires :

$$\theta_1 = \cdots = \theta_J \Leftrightarrow \begin{cases} \theta_2 - \theta_1 = 0; \\ \theta_3 - \theta_1 = 0; \\ \vdots \\ \theta_J - \theta_1 = 0 \end{cases}$$

Ce test est donc équivalent au test de Wald de la nullité de $C\theta$ où

$$C = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Dans ce cas-là, la statistique du test de Wald a une forme bien particulière que l'on va expliciter

Néanmoins, il se trouve que la statistique du test de Wald peut être mise sous une forme bien particulière qui s'interprète bien. Afin de voir apparaître ces quantités "bien interprétables" on va voir une façon alternative d'arriver au test de Wald. Pour cela on commence par définir les quantités suivantes :

Définition 36

- *On appelle* ***somme des carrées*** *la quantité :*

$$SC = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{i,j} - \overline{Y})^2 = \|Y - \overline{Y} 1_n\|_2^2$$

- *On appelle* ***somme des carrées intraclasse*** *la quantité :*

$$SC_{intra} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{i,j} - \overline{Y_{\bullet,j}})^2 = \|Y - \hat{Y}\|_2^2$$

- *On appelle* ***somme des carrées interclasse*** *la quantité :*

$$SC_{inter} = \sum_{j=1}^{J} n_j (\overline{Y_{\bullet,j}} - \overline{Y})^2 = \|\hat{Y} - \overline{Y} 1_n\|_2^2$$

Remarque: On a vu que la somme des carrées intraclasse correspond aussi à l'inertie des résidus et la somme des carrées interclasse correspond aussi à l'inertie expliquée par le modèle linéaire.

On a aussi déjà montré que ces quantités vérifient :

Proposition 37

$$SC = SC_{intra} + SC_{inter}$$

On a aussi vu que chacune de ces quantités s'écrit comme la norme d'une projection de Y. En exploitant cette propriété on obtient le résultat suivant.

Théorème 38

Sous l'hypothèse $\mathcal{H}_0$ les variables aléatoires SC_{intra} et SC_{inter} sont indépendantes de loi

$$\frac{SC_{intra}}{\sigma^2} \sim \chi^2(n-J) \text{ et } \frac{SC_{inter}}{\sigma^2} \sim \chi^2(J-1).$$

Démonstration : On a déjà montré que si on pose E l'espace vectoriel engendré par les colonnes de X (et qui contient donc $\text{Vect}(1_n)$) et G le complément orthogonal de $\text{Vect}(1_n)$ dans E alors

$$SC_{intra} = \|P_{E^\perp}(Y)\|_2^2 \text{ et } SC_{inter} = \|P_G(Y)\|_2^2.$$

On a aussi vu que $X\theta \in E$ donc

$$P_{E^\perp}(Y) = P_{E^\perp}(X\theta + \varepsilon) = P_{E^\perp}(\varepsilon).$$

Maintenant, sous $\mathcal{H}_0$ on a que les θ_i sont constant et donc $\theta \in \text{Vect}(1_n)$. Or, on vérifie facilement que $X\theta$ est aussi un vecteur constant donc $X\theta \in \text{Vect}(1_n)$ d'où

$$P_G(Y) = P_G(X\theta + \varepsilon) = P_G(\varepsilon).$$

Comme $E^\perp$ et G sont orthogonaux de dimensions respectives $n-J$ et $J-1$ et $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ alors par le théorème de Cochran on en déduit que $P_{E^\perp}(\varepsilon)$ et $P_G(\varepsilon)$ sont indépendants et vérifient

$$\frac{\|P_{E^\perp}(\varepsilon)\|_2^2}{\sigma^2} \sim \chi^2(n-J) \text{ et } \frac{\|P_G(\varepsilon)\|_2^2}{\sigma^2} \sim \chi^2(J-1),$$

d'où le résultat. ■

Remarque: En conséquence des deux résultats précédents on en déduit que $\frac{SC}{\sigma^2} \sim \chi^2(n-1)$.

Corollaire 39

Sous l'hypothèse $\mathcal{H}_0$, la statistique

$$F = \frac{SC_{inter}/(J-1)}{SC_{intra}/(N-J)}$$

suit la loi $\mathcal{F}(J-1, n-J)$.

On remarque que sous l'hypothèse $\mathcal{H}_1$ alors il y aura au moins un i pour lequel $\overline{Y_{i,\bullet}} \underset{n\to+\infty}{\not\to} \overline{Y}$ et donc $SC_{inter} \to \infty$ et $F \to \infty$. On va donc rejeter l'hypothèse $\mathcal{H}_0$ au risque α lorsque F est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(J-1, n-J)$.

On résume en général le résultat d'ANOVA dans le tableau suivant :

	Somme des carrées (SC)	Degrés de liberté (DDL)	Rapports SC/DDL	Résultat
Interclasse	SC_{inter}	$J-1$	$A = \frac{SC_{inter}}{J-1}$	$F = \frac{A}{B}$
Intraclasse	SC_{intra}	$n-J$	$B = \frac{SC_{intra}}{n-J}$	
Total	SC	$n-1$	$\frac{SC}{n-1}$	

On vérifie maintenant l'équivalence avec le test de Wald.

Théorème 40

Le test de Wald pour la matrice

$$C = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

a pour statistique

$$F = \frac{SC_{inter}/(J-1)}{SC_{intra}/(N-J)}.$$

Démonstration : Comme C est de taille $(J-1)\times J$ alors la statistique du test de Wald s'écrit

$$\frac{\langle C\hat{\theta}, (C\Sigma {}^tC)^{-1}C\hat{\theta}\rangle}{(J-1)\widetilde{\sigma}^2} \text{ où } \Sigma = ({}^tXX)^{-1}.$$

De plus, on a vu que $\widetilde{\sigma}^2 = \frac{1}{N-J}\|Y-\hat{Y}\|^2 = \frac{1}{N-J}SC_{intra}$. Du coup, il ne reste plus qu'à vérifier que $\langle C\hat{\theta}, (C\Sigma {}^tC)^{-1}C\hat{\theta}\rangle = SC_{inter}$. Comme $\hat{\theta}_j = \overline{Y_{\bullet,j}}$ on en déduit que

$$n\overline{Y} = \sum_{j=1}^{J} n_j \overline{Y_{\bullet,j}}$$

d'où

$$\begin{aligned}
SC_{inter} &= \sum_{i=1}^{J} n_i(\overline{Y_{\bullet,i}} - \overline{Y})^2 \\
&= \sum_{i=1}^{J} n_i \left(\hat{\theta}_i - \sum_{j=1}^{J}\frac{n_j}{n}\hat{\theta}_j\right)^2 \\
&= \sum_{i=1}^{J} n_i\hat{\theta}_i^2 - \sum_{i=1}^{J} n_i\hat{\theta}_i \times 2\sum_{j=1}^{J}\frac{n_j}{n}\hat{\theta}_j + \sum_{i=1}^{J} n_i\left(\sum_{j=1}^{J}\frac{n_j}{n}\hat{\theta}_j\right)^2 \\
&= \sum_{i=1}^{J} n_i\hat{\theta}_i^2 - 2\sum_{i,j=1}^{J}\frac{n_in_j}{n}\hat{\theta}_i\hat{\theta}_j + n\left(\sum_{j=1}^{J}\frac{n_j}{n}\hat{\theta}_j\right)^2 \\
&= \sum_{i=1}^{J} n_i\hat{\theta}_i^2 - 2\sum_{i,j=1}^{J}\frac{n_in_j}{n}\hat{\theta}_i\hat{\theta}_j + \sum_{i,j=1}^{J}\frac{n_in_j}{n}\hat{\theta}_i\hat{\theta}_j \\
&= \sum_{i=1}^{J} n_i\hat{\theta}_i^2 - \sum_{i,j=1}^{J}\frac{n_in_j}{n}\hat{\theta}_i\hat{\theta}_j \\
&= \langle\hat{\theta}, D\hat{\theta}\rangle - \frac{1}{n}\langle\hat{\theta}, M\hat{\theta}\rangle,
\end{aligned}$$

où D est la matrice diagonale des n_i et $M_{i,j} = n_in_j$. Maintenant, on peut écrire

$$\langle C\hat{\theta}, (C\Sigma {}^tC)^{-1}C\hat{\theta}\rangle = \langle\hat{\theta}, {}^tC(C\Sigma {}^tC)^{-1}C\hat{\theta}\rangle$$

donc il suffit de prouver que ${}^tC(C\Sigma {}^tC)^{-1}C = D - \frac{1}{n}M$ pour conclure. On a déjà vu que $\Sigma = ({}^tXX)^{-1}$ est la matrice diagonale des $1/n_i$ donc

$$C\Sigma = \begin{pmatrix} -\frac{1}{n_1} & \frac{1}{n_2} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\frac{1}{n_1} & 0 & \cdots & 0 & \frac{1}{n_J} \end{pmatrix} \Rightarrow C\Sigma {}^tC = \begin{pmatrix} \frac{1}{n_1}+\frac{1}{n_2} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \frac{1}{n_1} & \frac{1}{n_1}+\frac{1}{n_3} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{n_1} \\ \frac{1}{n_1} & \cdots & \frac{1}{n_1} & \frac{1}{n_1}+\frac{1}{n_J} \end{pmatrix}$$

c'est à dire

$$C\Sigma {}^tC = \frac{1}{n_1}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} + \begin{pmatrix} \frac{1}{n_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_3} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{n_J} \end{pmatrix}.$$

Ensuite, on pose D' la matrice diagonale de $n_2, \cdots, n_J$ et M' tel que $M'_{i,j} = n_{i+1}n_{j+1}$ alors, en utilisant le fait que $n_2 + \cdots + n_J = n - n_1$ on obtient

$$\left(D' - \frac{1}{n}M'\right)(C\Sigma^t C) = \left(\frac{1}{n_1}\begin{pmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots\\ 1 & \cdots & 1\end{pmatrix} + \begin{pmatrix}\frac{1}{n_2} & & 0\\ & \ddots & \\ 0 & & \frac{1}{n_J}\end{pmatrix}\right)\left(\begin{pmatrix}n_2 & & 0\\ & \ddots & \\ 0 & & n_J\end{pmatrix} - \frac{1}{n}\begin{pmatrix}n_2^2 & \cdots & n_2n_J\\ \vdots & \ddots & \vdots\\ n_Jn_2 & \cdots & n_J^2\end{pmatrix}\right) = \frac{1}{n_1}\begin{pmatrix}n_2 & \cdots & n_J\\ \vdots & & \vdots\\ n_2 & \cdots & n_J\end{pmatrix} - \frac{n-n_1}{nn_1}\begin{pmatrix}n_2 & \cdots & n_J\\ \vdots & & \vdots\\ n_2 & \cdots & n_J\end{pmatrix} + I_{J-1}$$
$$- \frac{1}{n}\begin{pmatrix}n_2 & \cdots & n_J\\ \vdots & & \vdots\\ n_2 & \cdots & n_J\end{pmatrix} = I_{J-1}.$$

Donc $(C\Sigma^t C)^{-1} = D' - \frac{1}{n}M'$. Maintenant,

$$(C\Sigma^t C)^{-1}C = D'C - \frac{1}{n}M'C = \begin{pmatrix}-n_2 & n_2 & 0 & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 0\\ -n_J & 0 & \cdots & 0 & n_J\end{pmatrix} - \frac{1}{n}\begin{pmatrix}-n_2(n-n_1) & n_2^2 & \cdots & n_2n_J\\ \vdots & \vdots & \ddots & \vdots\\ -n_J(n-n_1) & n_Jn_2 & \cdots & n_J^2\end{pmatrix}$$

donc ${}^tC(C\Sigma^t C)^{-1}C$ est égal à

$$\begin{pmatrix}n-n_1 & -n_2 & \cdots & \cdots & -n_J\\ -n_2 & n_2 & 0 & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 0\\ -n_J & 0 & \cdots & 0 & n_J\end{pmatrix} - \frac{1}{n}\begin{pmatrix}(n-n_1)^2 & -n_2(n-n_1) & \cdots & -n_J(n-n_1)\\ -n_2(n-n_1) & n_2^2 & \cdots & n_2n_J\\ \vdots & \vdots & \ddots & \vdots\\ -n_J(n-n_1) & n_Jn_2 & \cdots & n_J^2\end{pmatrix}$$
$$= \begin{pmatrix}0 & 0 & \cdots & \cdots & 0\\ 0 & n_2 & 0 & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 0\\ 0 & 0 & \cdots & 0 & n_J\end{pmatrix} - \frac{1}{n}\begin{pmatrix}-n_1(n-n_1) & n_2n_1 & \cdots & n_Jn_1\\ n_2n_1 & n_2^2 & \cdots & n_2n_J\\ \vdots & \vdots & \ddots & \vdots\\ n_Jn_1 & n_Jn_2 & \cdots & n_J^2\end{pmatrix}$$
$$= \begin{pmatrix}n_1 & & 0\\ & \ddots & \\ 0 & & n_J\end{pmatrix} - \frac{1}{n}\begin{pmatrix}n_1^2 & \cdots & n_Jn_1\\ \vdots & \ddots & \vdots\\ n_Jn_1 & \cdots & n_J^2\end{pmatrix},$$

ce qui conclue la preuve. ■

4 Exemple d'application sur R

On considère les données de l'article : Ilvonen, J.J. and Jukka, S., *Phylogeny affects host's weight, immune response and parasitism in damselflies and dragonflies* (2016), Royal Society open science, **3**, 160421.
contenant la longueur (en mm) des ailes de 486 libellules, mâles et femelles, de 19 espèces différentes.
On se demande s'il y a une différence significative de longueur des ailes selon le sexe et l'espèce.
On peut représenter visuellement l'effet d'une variable qualitative sur une variable quantitative

Espece	Sexe	Aile
Aeshna grandis	M	45.10
Aeshna grandis	M	47.41
Aeshna grandis	M	46.79
Aeshna grandis	M	46.07
Aeshna grandis	M	45.88
Aeshna grandis	M	44.10
⋮	⋮	⋮

en regardant le boxplot de la variable quantitative séparée par les modalités de la variable qualitative.

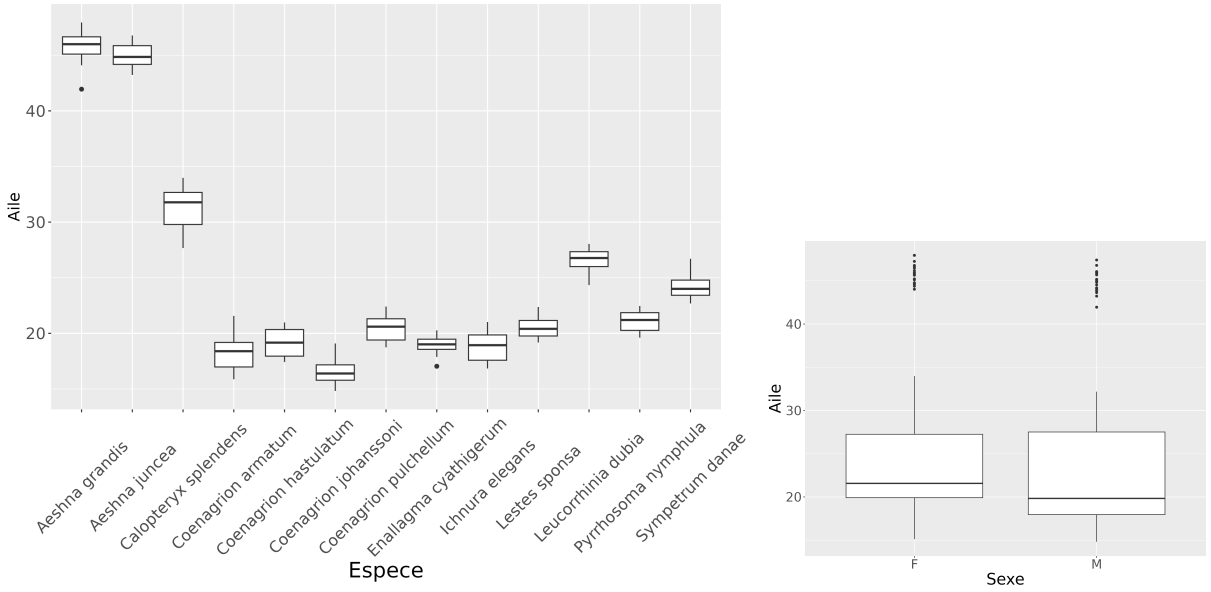


FIGURE 1.1 – Diagramme en boite des longueurs d'aile des libellules séparés par espèce (à gauche) ou par sexe (à droite)

Visuellement on peut observer des grosses différences entre les espèces mais pas vraiment entre les sexes. On cherche alors à tester ça en utilisant le modèle ANOVA. Avec la fonction *aov* de R on obtient les tables d'ANOVA suivantes qui confirment ce qu'on observe visuellement.

```
            Df Sum Sq Mean Sq F value Pr(>F)
Especce     12  25440  2120.0    1478 <2e-16 ***
Residuals  273    392     1.4
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Sexe         1    133  133.12   1.471  0.226
Residuals  284  25699   90.49
```

Si on utilise la fonction *lm* pour faire le modèle alors on obtient le résultat suivant selon qu'on utilise un intercept ou pas.

```
Call:
lm(formula = Aile ~ Espece, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9364 -0.9378  0.0330  0.8776  3.2223

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  45.8864     0.2554 179.697  < 2e-16 ***
EspeceAeshna juncea          -0.9632     0.3611  -2.667  0.00811 **
EspeceCalopteryx splendens  -14.6177     0.3611 -40.478  < 2e-16 ***
EspeceCoenagrion armatum    -27.5386     0.3611 -76.258  < 2e-16 ***
EspeceCoenagrion hastulatum -26.7700     0.3611 -74.129  < 2e-16 ***
EspeceCoenagrion johanssoni -29.3982     0.3611 -81.407  < 2e-16 ***
EspeceCoenagrion pulchellum -25.4514     0.3611 -70.478  < 2e-16 ***
EspeceEnallagma cyathigerum -26.8755     0.3611 -74.421  < 2e-16 ***
EspeceIchnura elegans       -26.9800     0.3611 -74.711  < 2e-16 ***
EspeceLestes sponsa         -25.4286     0.3611 -70.415  < 2e-16 ***
EspeceLeucorrhinia dubia    -19.2900     0.3611 -53.416  < 2e-16 ***
EspecePyrrhosoma nymphula   -24.7700     0.3611 -68.591  < 2e-16 ***
EspeceSympetrum danae       -21.7259     0.3611 -60.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.198 on 273 degrees of freedom
Multiple R-squared:  0.9848,    Adjusted R-squared:  0.9842
F-statistic:  1478 on 12 and 273 DF,  p-value: < 2.2e-16
```

(a) Avec intercept

```
Call:
lm(formula = Aile ~ Espece - 1, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9364 -0.9378  0.0330  0.8776  3.2223

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
EspeceAeshna grandis        45.8864     0.2554  179.70  <2e-16 ***
EspeceAeshna juncea         44.9232     0.2554  175.93  <2e-16 ***
EspeceCalopteryx splendens  31.2686     0.2554  122.45  <2e-16 ***
EspeceCoenagrion armatum    18.3477     0.2554   71.85  <2e-16 ***
EspeceCoenagrion hastulatum 19.1164     0.2554   74.86  <2e-16 ***
EspeceCoenagrion johanssoni 16.4882     0.2554   64.57  <2e-16 ***
EspeceCoenagrion pulchellum 20.4350     0.2554   80.03  <2e-16 ***
EspeceEnallagma cyathigerum 19.0109     0.2554   74.45  <2e-16 ***
EspeceIchnura elegans       18.9064     0.2554   74.04  <2e-16 ***
EspeceLestes sponsa         20.4577     0.2554   80.11  <2e-16 ***
EspeceLeucorrhinia dubia    26.5964     0.2554  104.16  <2e-16 ***
EspecePyrrhosoma nymphula   21.1164     0.2554   82.69  <2e-16 ***
EspeceSympetrum danae       24.1605     0.2554   94.61  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.198 on 273 degrees of freedom
Multiple R-squared:  0.9981,    Adjusted R-squared:  0.998
F-statistic: 1.105e+04 on 13 and 273 DF,  p-value: < 2.2e-16
```

(b) Sans intercept

On peut voir que si le modèle a un intercept alors la variable associée à la modalité *Aesha juncea* disparaît car elle est considérée comme modalité de référence. Les coefficients sont alors modifiés comme expliqué précédemment.

V Analyse de la variance à deux facteurs

1 Écriture du modèle

On considère maintenant que l'on a deux variables explicatives qualitatives avec J et K modalités respectivement. De façon similaire à la section précédente, on note $y_{i,j,k}$ la valeur de la variable d'intérêt Y pour le i-ième individu possédant la j-ème modalité de la première variable explicative et la k-ème modalité de la deuxième variable explicative.
⚠ On fait l'hypothèse qu'on a le même nombre d'individu, noté I, possédant chaque paire de modalité. On a donc $n = IJK$ individus au total.

Dans le modèle ANOVA à deux facteurs on suppose que $y_{i,j,k}$ est issu d'une variable

$$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon_{i,j,k}$$

où les $\varepsilon_{i,j,k}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Les paramètres du modèles sont les suivants :

- μ : La moyenne générale (l'intercept)
- α_j : L'effet différentiel (en écart à la moyenne) de la j-ème modalité de la première variable.
- β_k : L'effet différentiel (en écart à la moyenne) de la k-ème modalité de la seconde variable.
- $\gamma_{j,k}$: L'effet d'interaction entre la j-ème modalité de la première variable et la k-ème modalité de la seconde variable.

Ces paramètres doivent vérifier les contraintes suivantes :

$$\sum_{j=1}^{J} \alpha_j = 0, \sum_{k=1}^{K} \beta_k = 0, \sum_{j=1}^{J} \gamma_{j,k} = 0 \text{ pour tout } k \text{ et } \sum_{k=1}^{K} \gamma_{j,k} = 0 \text{ pour tout } j.$$

Définition 41

On note :

- $\overline{Y}$ *: la moyenne de la variable d'intérêt.*
- $\overline{Y_{\bullet,j,\bullet}}$ *: la moyenne de la variable d'intérêt pour les individus possédant la j-ème modalité de la première variable.*
- $\overline{Y_{\bullet,\bullet,k}}$ *: la moyenne de la variable d'intérêt pour les individus possédant la k-ème modalité de la seconde variable.*
- $\overline{Y_{\bullet,j,k}}$ *: la moyenne de la variable d'intérêt pour les individus possédant la j-ème modalité de la première variable et la k-ème modalité de la seconde variable.*

Les estimateurs des paramètres du modèle sont alors

$$\hat{\mu} = \overline{Y}, \quad \hat{\alpha}_j = \overline{Y_{\bullet,j,\bullet}} - \hat{\mu}, \quad \hat{\beta}_k = \overline{Y_{\bullet,\bullet,k}} - \hat{\mu}$$

et

$$\hat{\gamma}_{j,k} = \overline{Y_{\bullet,j,k}} - \hat{\alpha}_j - \hat{\beta}_k - \hat{\mu} = \overline{Y_{\bullet,j,k}} - \overline{Y_{\bullet,j,\bullet}} - \overline{Y_{\bullet,\bullet,k}} + \overline{Y}.$$

2 Tests statistiques

On considère les trois hypothèses suivantes :

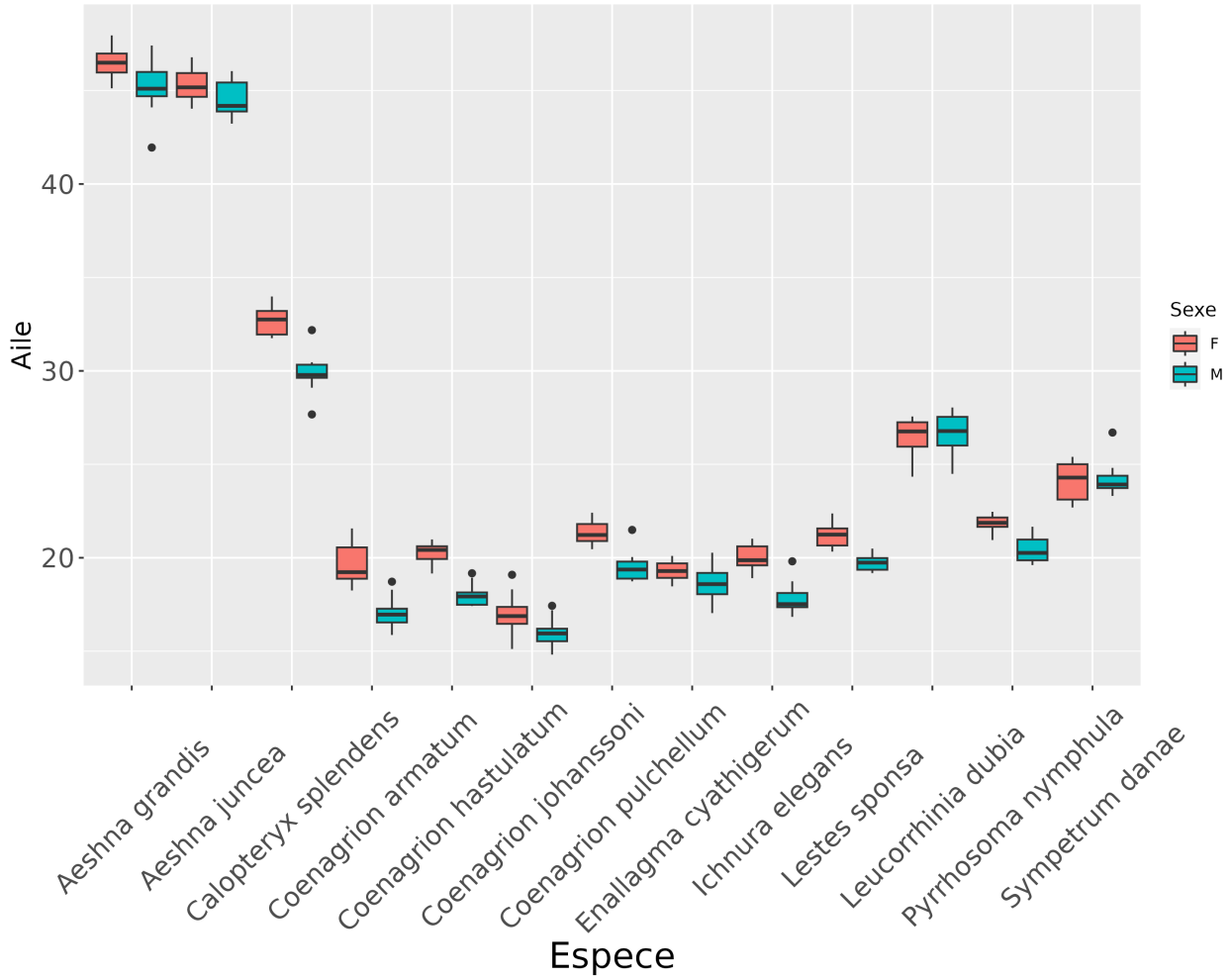
- $\mathcal{H}_0^A : \{\alpha_1 = \cdots = \alpha_J = 0\}$, c'est à dire que la première variable qualitative n'a pas d'effet sur Y.
- $\mathcal{H}_0^B : \{\beta_1 = \cdots = \beta_K = 0\}$, c'est à dire que la deuxième variable qualitative n'a pas d'effet sur Y.
- $\mathcal{H}_0^C : \{\gamma_{j,k} = 0$ pour tout $j, k\}$, c'est à dire qu'il n'y a pas d'effet d'interaction entre les variables.

Afin de faire les tests, on construit le tableau suivant :

	SC	DDL	$\frac{\text{SC}}{\text{DDL}}$	Résultat
Premier	$IK\sum_{j=1}^{J}(\overline{Y_{\bullet,j,\bullet}} - \overline{Y})^2$	$J-1$	A	$F_A = \frac{A}{D}$
Deuxième	$IJ\sum_{k=1}^{K}(\overline{Y_{\bullet,\bullet,k}} - \overline{Y})^2$	$K-1$	B	$F_B = \frac{B}{D}$
Intéractions	$I\sum_{j,k=1}^{J,K}(\overline{Y_{\bullet,j,k}} - \overline{Y_{\bullet,j,\bullet}} - \overline{Y_{\bullet,\bullet,k}} + \overline{Y})^2$	$(J-1)(K-1)$	C	$F_C = \frac{C}{D}$
Résidus	$\sum_{i,j,k=1}^{I,J,K}(Y_{i,j,k} - \overline{Y_{\bullet,j,k}})^2$	$(I-1)JK$	D	
Total	$\sum_{i,j,k=1}^{I,J,K}(Y_{i,j,k} - \overline{Y})^2$	$IJK-1$		

- On rejette $\mathcal{H}_0^A$ au risque α si F_A est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(J-1,\ (I-1)JK)$.
- On rejette $\mathcal{H}_0^B$ au risque α si F_B est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(K-1,\ (I-1)JK)$.
- On rejette $\mathcal{H}_0^C$ au risque α si F_C est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}((J-1)(K-1),\ (I-1)JK)$.

Exemple : On reprend l'exemple des libellules. On commence par visualiser le boxplot des longueurs d'aile de libellules séparé par espèce et par sexe en même temps.



On peut observer que contrairement à ce qu'on observe lorsqu'on regardait juste l'influence du sexe seul sur les longueurs d'aile, le fait de prendre en compte les différence entre espèce à l'air de faire apparaître une différence entre les sexes. On peut voir que les femelles ont tendance à avoir des ailes au moins aussi grandes, voir même plus grande, que les mâles. Néanmoins, cette différence n'est pas toujours la même selon l'espèce ce qui montre l'existence d'un effet d'interaction entre le sexe et l'espèce. Appliquer un test ANOVA à deux facteurs sur ces données démontre bien l'existence de ces effets.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Espece	12	25440	2120.0	2777.953	< 2e-16 ***
Sexe	1	133	133.1	174.431	< 2e-16 ***
Espece:Sexe	12	60	5.0	6.561	3.11e-10 ***
Residuals	260	198	0.8		

VI Sélection de modèle

1 Présentation du problème

On considère que l'on a p variables quantitatives. On appelle **modèle** m un sous-ensemble de $\{1, \cdots, p\}$ correspondant à la sélection d'un sous-ensemble des variables quantitatives et on

définit $\mathcal{M} \subset \mathcal{P}(\{1, \cdots, p\})$ un ensemble de modèle. La régression linéaire pour un modèle m s'écrit

$$Y_i = \theta_0 + \sum_{j \in m} \theta_j x_{i,j} + \varepsilon_i$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et **les θ_j sont non-nuls**. Le modèle s'écrit vectoriellement :

$$Y = X_e^{(m)} \theta^{(m)} + \varepsilon.$$

On notera $|m|$ le cardinal de m et $\hat{\theta}^{(m)}$, $\hat{\sigma}^{(m)}$ et $\tilde{\sigma}^{(m)}$ les estimateurs des paramètres pour le modèle m. On considère qu'il existe un vrai modèle $m_* \in \mathcal{M}$ tel que

$$Y = X_e^{(m_*)} \theta^{(m_*)} + \varepsilon.$$

Remarque: Si on souhaite comparer deux modèles m_1 et m_2 imbriqués (c'est à dire que $m_1 \subset m_2$) alors cela revient à tester la nullité des θ_i pour $i \in m_2 \backslash m_1$ et on se ramène au test de Wald correspondant. Ce test est aussi souvent appelé ANOVA (notamment sur R) vu que le test de Wald fonctionne de façon similaire à un test ANOVA.

2 Divers critères de sélection de modèle

a Le critère CP de Mallow

Pour un modèle $m \in \mathcal{M}$, on note $\hat{Y}^{(m)} = X_e^{(m)} \theta^{(m)}$ les valeurs des Y_i estimées si on suppose que Y suit le modèle m. On commence à regarder le comportement de la norme du vecteur des résidus dans ce cas-là. Pour cela on pose $P^{(m)} = I_n - X_e^{(m)} ({}^t X_e^{(m)} X_e^{(m)})^{-1} X_e^{(m)}$ et $\mu = \mathbb{E}[Y] = X_e^{(m^*)} \theta^{(m_*)}$.

Proposition 42

$$\mathbb{E}\left[\|Y - \hat{Y}^{(m)}\|_2^2\right] = \langle \mu, P^{(m)} \mu \rangle + (n - (|m| + 1))\sigma^2.$$

Démonstration : On utilisera le lemme suivant :

Lemme 43

Soit X un vecteur aléatoire de $\mathbb{R}^d$ d'espérance v et de matrice de variance-covariance Σ. Soit $M \in \mathcal{M}_d(\mathbb{R})$ alors

$$\mathbb{E}[\langle X, MX \rangle] = \langle v, Mv \rangle + Tr(\Sigma M).$$

Démonstration : On a $\langle X, MX \rangle = \sum_{i=1}^n \sum_{j=1}^n X_i X_j M_{i,j}$ et $\mathbb{E}[X_i X_j] = \text{cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j] = \Sigma_{j,i} + v_i v_j$ donc

$$\mathbb{E}[\langle X, MX \rangle] = \sum_{i=1}^n \sum_{j=1}^n v_i v_j M_{i,j} + \sum_{i=1}^n \sum_{j=1}^n \Sigma_{j,i} M_{i,j} = \langle v, Mv \rangle + \text{Tr}(\Sigma M).$$

■

On a $Y - \hat{Y}^{(m)} = Y - X_e^{(m)} \hat{\theta}^{(m)} = P^{(m)} Y$. Or, comme $P^{(m)}$ est une matrice symétrique de projection alors

$$\|Y - \hat{Y}^{(m)}\|_2^2 = \langle P^{(m)} Y, P^{(m)} Y \rangle = \langle Y, P^{(m)} P^{(m)} Y \rangle = \langle Y, P^{(m)} Y \rangle$$

donc le lemme précédant nous donne

$$\mathbb{E}\left[\|Y - \hat{Y}^{(m)}\|_2^2\right] = \langle \mu, P^{(m)} \mu \rangle + \text{Tr}(\sigma^2 P^{(m)}).$$

Or, $\mathrm{Tr}(I_n) = n$ et

$$\mathrm{Tr}(X_e^{(m)}({}^tX_e^{(m)}X_e^{(m)})^{-1t}X_e^{(m)}) = \mathrm{Tr}(({}^tX_e^{(m)}X_e^{(m)})^{-1t}X_e^{(m)}X_e^{(m)}) = \mathrm{Tr}(I_{|m|+1}) = |m|+1,$$

d'où le résultat. ■

On remarque que, même si le modèle est mauvais, le fait d'augment le nombre de variables explicatives (et donc $|m|$) fait diminuer $(n-(|m|+1))\sigma^2$ et donc la norme des résidus. Afin d'avoir un critère plus fiable de la qualité du modèle on va plutôt regarder la qualité des prédictions du modèles sur des nouvelles données Y' générées avec le même vrai modèle m^* (et donc les mêmes variables explicatives) mais avec un nouvel aléa :

$$Y' = X_e^{(m_*)}\theta^{(m_*)} + \varepsilon',$$

où $\varepsilon' \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ avec ε' indépendant de ε. En utilisant le modèle m on estime alors Y' par $X_e^{(m)}\hat{\theta}^{(m)} = \hat{Y}^{(m)}$ mais dans ce cas-là l'erreur d'estimation devient :

Proposition 44

$$\mathbb{E}\left[\|Y' - \hat{Y}^{(m)}\|_2^2\right] = \langle \mu, P^{(m)}\mu\rangle + (n + (|m|+1))\sigma^2.$$

Démonstration : On utilisera le lemme suivant :

Lemme 45

Soient X, Y deux vecteurs aléatoires indépendants de $\mathbb{R}^d$ de vecteurs espérances v et v' et de matrices de variance-covariance Σ et Σ'. Alors

$$\mathbb{E}\left[\|X - Y\|_2^2\right] = Tr(\Sigma) + Tr(\Sigma') + \|v - v'\|_2^2.$$

Démonstration : On a $\|X-Y\|_2^2 = \sum_{i=1}^n (X_i - Y_i)^2 = \sum_{i=1}^n (X_i^2 + Y_i^2 - 2X_iY_i)$. Comme X et Y sont indépendants alors $\mathbb{E}[X_iY_i] = \mathbb{E}[X_i]\mathbb{E}[Y_i]$ donc

$$\begin{aligned}
\mathbb{E}[\|X-Y\|_2^2] &= \sum_{i=1}^n \left(\mathbb{E}\left[X_i^2\right] + \mathbb{E}\left[Y_i^2\right] - 2\mathbb{E}[X_i]\mathbb{E}[Y_i]\right) \\
&= \sum_{i=1}^n \left(\mathrm{Var}(X_i) + \mathbb{E}[X_i]^2 + \mathrm{Var}(Y_i) + \mathbb{E}[Y_i]^2 - 2\mathbb{E}[X_i]\mathbb{E}[Y_i]\right) \\
&= \sum_{i=1}^n \left(\mathrm{Var}(X_i) + \mathrm{Var}(Y_i) + (\mathbb{E}[X_i] - \mathbb{E}[Y_i])^2\right) \\
&= \sum_{i=1}^n \left(\Sigma_{i,i} + \Sigma'_{i,i} + (v_i - v'_i)^2\right) \\
&= \mathrm{Tr}(\Sigma) + \mathrm{Tr}(\Sigma') + \|v - v'\|_2^2.
\end{aligned}$$

■

Maintenant, on peut directement utiliser ce lemme car Y' dépend seulement de ε' et $\hat{Y}^{(m)}$ dépend seulement de ε qui sont bien indépendant. On a $Y' = X_e^{(m_*)}\theta^{(m_*)} + \varepsilon'$ donc $\mathbb{E}[Y'] = X_e^{(m_*)}\theta^{(m_*)} = \mu$ et $\mathrm{Var}(Y') = \sigma^2 I_n$. De plus, on a $\hat{Y}^{(m)} = (I_n - P^{(m)})Y$ donc $\mathbb{E}[\hat{Y}^{(m)}] = (I_n - P^{(m)})\mu$ et $\mathrm{Var}(\hat{Y}^{(m)}) = \sigma^2(I_n - P^{(m)})$ car $I_n - P^{(m)}$ est orthogonale. Cela donne alors

$$\begin{aligned}
\mathbb{E}\left[\|Y - X\|_2^2\right] &= \mathrm{Tr}(\sigma^2(I_n)) + \mathrm{Tr}(\sigma^2(I_n - P^{(m)})) + \|\mu - (I_n - P^{(m)})\mu\|_2^2 \\
&= n\sigma^2 + n\sigma^2 - (n - (|m|+1))\sigma^2 + \|P^{(m)}\mu\|_2^2 \\
&= (n + (|m|+1))\sigma^2 + \langle P^{(m)}\mu, P^{(m)}\mu\rangle \\
&= (n + (|m|+1))\sigma^2 + \langle \mu, P^{(m)}\mu\rangle.
\end{aligned}$$

■

Minimiser les erreurs de prédiction du modèle sur de nouvelles valeurs revient alors à minimiser la quantité

$$\frac{\mathbb{E}\left[\|Y' - \hat{Y}^{(m)}\|_2^2\right]}{n\sigma^2} = \mathbb{E}\left[\|Y - \hat{Y}^{(m)}\|_2^2\right] + 2(|m|+1)\sigma^2 = \frac{\mathbb{E}\left[\left(\hat{\sigma}^{(m)}\right)^2\right]}{\sigma^2} + \frac{2(|m|+1)}{n}.$$

Si on estime $\mathbb{E}\left[\left(\hat{\sigma}^{(m)}\right)^2\right]$ par $\left(\hat{\sigma}^{(m)}\right)^2$ et σ^2 par $\left(\hat{\sigma}^{(m_c)}\right)^2$, l'estimation de σ^2 utilisant le modèle complet $m_c = \{1, \cdots, p\}$ alors on appelle **critère C_p de Mallow** la quantité

$$C_p(m) = \left(\frac{\hat{\sigma}^{(m)}}{\hat{\sigma}^{(m_c)}}\right)^2 + \frac{2(|m|+1)}{n}.$$

et on sélectionne le modèle m qui minimise $C_p(m)$.
Remarque: Si on suppose σ^2 connu alors on n'a pas besoin de l'estimer et le critère est simplement

$$C_p(m) = \left(\frac{\hat{\sigma}^{(m)}}{\sigma}\right)^2 + \frac{2(|m|+1)}{n}.$$

b Les critères AIC et BIC

Ces deux critères sont basés sur le comportement asymptotique de la divergence de Kullback-Leibler (un outil mesurant la similarité entre deux lois de probabilités) entre la loi de Y pour le modèle m^* et la loi de Y estimée pour le modèle m. Selon les hypothèses faites sur le modèle, on peut arriver à deux expressions asymptotiques différentes de cette divergence.

Définition 46

On considère un modèle pour n individus défini par un vecteur θ de k paramètres. On note L la vraisemblance du modèle et $\hat{\theta}$ l'EMV des paramètres.

- *On définit le* ***critère d'information d'Akaïke*** *(AIC) du modèle par*

$$AIC = -2\log(L(\hat{\theta})) + 2k$$

- *On définit le* ***critère d'information Bayésien*** *(BIC) du modèle par*

$$BIC = -2\log(L(\hat{\theta})) + k\log(n)$$

Remarque: Ces critères sont génériques et peuvent s'appliquer à la comparaison de n'importe quel types de modèles du temps qu'ils possèdent une vraisemblance.

Proposition 47

Dans le cas des modèles linéaires, pour un modèle m on a

$$AIC(m) = n(1 + \log(2\pi)) + 2n\log(\hat{\sigma}^{(m)}) + 2(|m|+2)$$

$$BIC(m) = n(1 + \log(2\pi)) + 2n\log(\hat{\sigma}^{(m)}) + \log(n)(|m|+2)$$

Démonstration : Pour un modèle m avec un intercept et $|m|$ variables explicatives on a $k = |m| + 2$ paramètres (θ_0, les θ_i pour $i \in m$ et σ^2). De plus, on a

$$\begin{aligned}\log\left(L(\hat{\theta}^{(m)}, \hat{\sigma}^{(m)})\right) &= \log\left(\frac{1}{(2\pi\left(\hat{\sigma}^{(m)}\right)^2)^{n/2}} \exp\left(-\frac{1}{2\left(\hat{\sigma}^{(m)}\right)^2}\|Y - \hat{Y}^{(m)}\|_2^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - n\log(\hat{\sigma}^{(m)}) - \frac{1}{2\left(\hat{\sigma}^{(m)}\right)^2}\|Y - \hat{Y}^{(m)}\|_2^2 \\ &= -\frac{n}{2}(1 + \log(2\pi)) - n\log(\hat{\sigma}^{(m)}),\end{aligned}$$

où on a utilisé le fait que $\|Y - \hat{Y}^{(m)}\|_2^2 = n\left(\hat{\sigma}^{(m)}\right)^2$ à la dernière ligne. ■

Remarque: Si on suppose σ^2 connu alors le modèle a seulement $|m| + 1$ paramètres et la log-vraisemblance s'écrit alors

$$\begin{aligned}\log\left(L(\hat{\theta}^{(m)}, \sigma^2)\right) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|Y - \hat{Y}^{(m)}\|_2^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|Y - \hat{Y}^{(m)}\|_2^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\left(\frac{\hat{\sigma}^{(m)}}{\sigma}\right)^2\end{aligned}$$

donc

$$\frac{AIC(m)}{n} = \log(2\pi) + \log(\sigma^2) + \left(\frac{\hat{\sigma}^{(m)}}{\sigma}\right)^2 + \frac{2(|m| + 1)}{n} = \text{Constante} + C_p(m).$$

Minimiser l'AIC est donc équivalent à minimiser le critère C_p de Mallow quand σ^2 est connu.

c La validation non-croisée

Une idée pour quantifier la performance d'un modèle est d'estimer l'erreur moyenne de ce modèle pour prédire de nouvelles valeurs. Pour cela, on sépare nos n individus en deux ensembles : un ensemble dit d'**entraînement** et un ensemble dit de **test**. En général la taille de l'ensemble d'entraînement est choisie comme étant plus grosse que la taille de l'ensemble de test (typiquement, 2/3 des individus et 1/3 des individus respectivement). Si on note $\hat{y}_i^{(\text{train})}$ l'estimation de y_i par le méthode ajusté sur les données d'entraînement alors on estime l'erreur quadratique de prédiction du modèle par

$$\frac{1}{\#\text{test}} \sum_{i \in \text{test}} \left(\hat{y}_i^{(\text{train})} - y_i\right)^2.$$

Remarques:

- Très facile à implémenter numériquement.
- Cette méthode est utilisée lorsqu'on a un grand nombre de données et ce n'est pas grave d'en "sacrifier" une partie que l'on met dans l'ensemble de test.
- Il faut aussi faire très attention à la façon dont les données sont séparées dans le groupe d'entraînement et de test. Un mauvais choix peut fausser les résultats.
- Cette méthode est appelé **holdout method** en anglais.

d La validation croisée d'un contre tous

Une idée similaire à la validation non-croisée mais qui est plus adaptée à des petits jeux de données est d'ajuster le modèle sur $n-1$ individu et calculer son erreur de prédiction sur l'individu restant mais de recommencer n fois de sorte que chaque individu ai été retiré une fois du jeu de données. Si on note $\hat{y}_i^{(-i)}$ l'estimation de y_i par le modèle ajusté sur tout les individus sauf le i-ème alors l'erreur quadratique de prédiction du modèle est estimée par

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i^{(-i)} - y_i\right)^2.$$

Remarques:

- Cette méthode est en général plus précise que la validation non-croisée mais elle est beaucoup plus lente car elle demande de faire n ajustements de modèle.
- Cette méthode est appelée **leave-one-out cross-validation** ou **LOOCV** en anglais.

3 Algorithme de recherche du meilleur modèle

La plupart du temps, l'ensemble des modèles considérés est $\mathcal{M} = \mathcal{P}(\{1, \cdots, p\})$ qui est de taille 2^p. Quand p est trop grand il est donc impossible de calculer la valeur des critères pour chaque modèle. Une méthode classique est d'utiliser un algorithme "stepwise" :

Algorithme 1 : Méthode ascendante de sélection de modèle

Entrées : Les données X, Y et un critère C à minimiser

1 Initialiser m à $\emptyset$.
2 Calculer $C(m)$ et $C(m \cup \{i\})$ pour tout $i \notin m$.
3 Choisir le modèle m' qui minimise le critère parmi ceux calculés à l'étape précédente.
4 **si** $m' = m$ **alors**
5 | Terminer l'algorithme et renvoyer m.
6 **sinon**
7 | Affecter m' à m et retourner à l'étape 2.

On peut aussi utiliser une méthode descendante qui commence à $m = \{1, \cdots, p\}$ et cherche à chaque étape quelle variable à retirer afin de minimiser le critère choisit. On peut aussi utiliser une méthode hybride qui commence à un modèle m quelconque et cherche à chaque étape quelle variable retirer ou rajouter afin de minimiser le critère choisit.

VII Transformation des variables

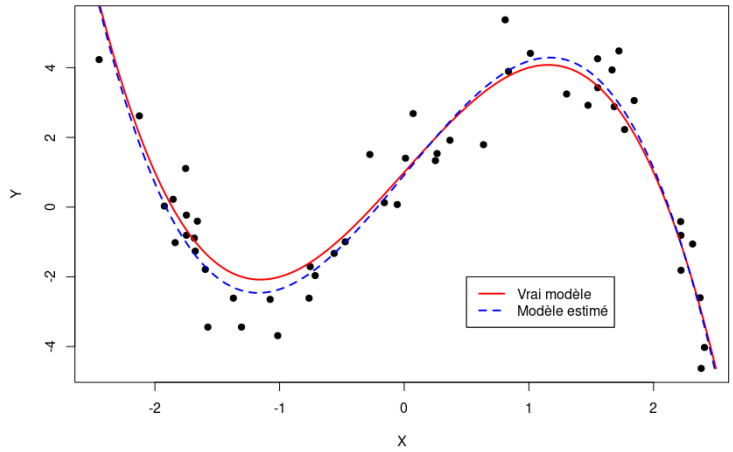
Ce n'est pas parce qu'un jeu de données contient des variables explicatives que l'on est obligé de toutes les utilisées telles qu'elles. On a vu avec la sélection de modèle comment choisir les bonnes variables mais on peut aussi modifier les variables du jeu de données.

1 La régression polynomiale

On considère avoir une seule variable explicative quantitative X. Le modèle de régression polynomiale de dimension d considère que les y_i sont des simulations de variables aléatoires

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_d x_i^d + \varepsilon_i.$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Ça correspond alors à un modèle linéaire pour les variables $X, X^2, \cdots, X^d$.



Exemple de régression polynomiale pour $d = 3$.

Comme il n'y a en général pas de raison de considérer pour un modèle de régression polynomial en dimension d que l'un des coefficients est nul, on choisit le degré de la régression polynomiale en calculant un des critère de sélection juste pour les modèles de régression polynomiale de dimension entre 0 et un degré maximal d_{max}.

```
[1] "d=1   AIC=96.708"
[1] "d=2   AIC=95.647"
[1] "d=3   AIC=-4.175"
[1] "d=4   AIC=-3.798"
[1] "d=5   AIC=-2.253"
[1] "d=6   AIC=-9.336"
[1] "d=7   AIC=-8.059"
[1] "d=8   AIC=-6.061"
[1] "d=9   AIC=-5.613"
[1] "d=10   AIC=-3.961"
[1] "d=11   AIC=-3.72"
[1] "d=12   AIC=-2.819"
[1] "d=13   AIC=-2.514"
[1] "d=14   AIC=-4.172"
[1] "d=15   AIC=-5.429"
[1] "d=16   AIC=-3.756"
[1] "d=17   AIC=-4.807"
[1] "d=18   AIC=-2.825"
[1] "d=19   AIC=-1.058"
[1] "d=20   AIC=-4.771"
```

```
[1] "d=1   BIC=101.918"
[1] "d=2   BIC=103.463"
[1] "d=3   BIC=6.246"
[1] "d=4   BIC=9.228"
[1] "d=5   BIC=13.378"
[1] "d=6   BIC=8.9"
[1] "d=7   BIC=12.783"
[1] "d=8   BIC=17.385"
[1] "d=9   BIC=20.439"
[1] "d=10   BIC=24.696"
[1] "d=11   BIC=27.542"
[1] "d=12   BIC=31.049"
[1] "d=13   BIC=33.959"
[1] "d=14   BIC=34.906"
[1] "d=15   BIC=36.254"
[1] "d=16   BIC=40.532"
[1] "d=17   BIC=42.086"
[1] "d=18   BIC=46.673"
[1] "d=19   BIC=51.045"
[1] "d=20   BIC=49.938"
```

Calcul des critères AIC et BIC pour les données de la figure précédente avec $d_{max} = 20$.

Remarque:

- On peut généraliser ce modèle à plusieurs variables en considérant des polynômes à plusieurs variables. Par exemple, si on a deux variables quantitative X et Z alors on peut écrire le modèle de régression polynomiale de dimensions 2 par

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 z_i + \theta_3 x_i^2 + \theta_4 x_i z_i + \theta_5 z_i^2 + \varepsilon_i.$$

- De façon plus générale on peut utiliser divers fonctions des données comme nouvelles variables. Par exemple :

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 z_i + \theta_3 \log(x_i) + \theta_4 e^{x_i + z_i} + \varepsilon_i$$

  est un modèle linéaire valide.

2 Le modèle ANCOVA

On considère maintenant que l'on a plusieurs variables explicatives quantitatives et qualitatives.

Exemple: On souhaite estimer la taille à l'âge adulte d'un enfant en fonction de son sexe et la taille de son père (qu'on note T). On considère alors le modèle :

$$Taille = \begin{cases} \theta_0 + \alpha_M + \theta_1 T + \beta_M T + \varepsilon_i, \text{ si sexe masculin;} \\ \theta_0 + \alpha_F + \theta_1 T + \beta_F T + \varepsilon_i, \text{ si sexe féminin;} \end{cases} \quad \text{avec } \alpha_M + \alpha_F = 0 \text{ et } \beta_M + \beta_F = 0.$$

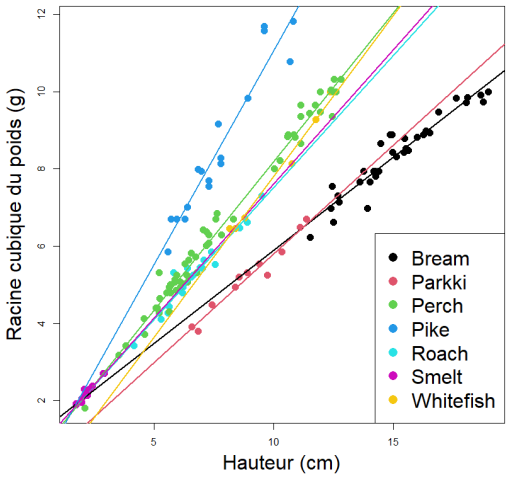
Ici, θ_1 nous donne l'effet de la taille du père sur l'ensemble de la population, α_M et α_F donne les différences de taille entre les hommes et les femmes et β_M et β_F donne la différence de l'effet de la taille du père sur la taille des hommes et des femmes. On est donc intéressé par le tests des hypothèses $\{\theta_1 = 0\}$, $\{\alpha_M = \alpha_F = 0\}$ et $\{\beta_M = \beta_F = 0\}$.

De façon générale, si on considère que l'on a p variables quantitatives $X^{(1)}, \cdots, X^{(p)}$ et q variables qualitatives. On note I_i le nombre de modalités de la i-ème variable qualitative et on note $F^{(i,j)}$ la variable qui vaut 1 si l'individu possède la j-ème modalité de la i-ème variable qualitative et 0 sinon. Le modèle ANCOVA (= analyse de la covariance) généralisé s'écrit alors

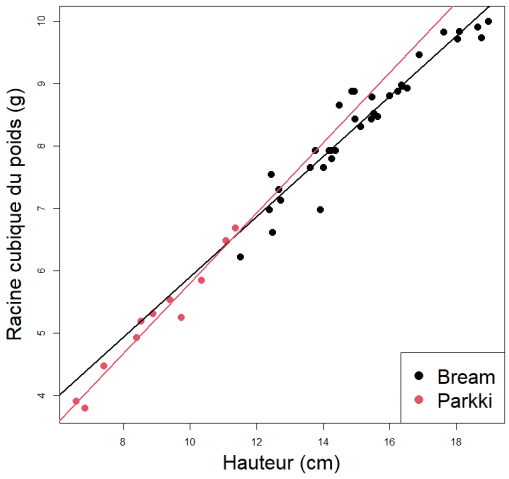
$$Y = \theta_0 + \sum_{i=1}^{p} \theta_i X^{(i)} + \sum_{j=1}^{q} \sum_{k=1}^{I_j} \mu_{j,k} F^{(j,k)} + \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{I_j} \mu_{i,j,k} X^{(i)} F^{(j,k)} + \sum_{j=1}^{q} \sum_{k=1}^{I_j} \sum_{j'=1}^{q} \sum_{k'=1}^{I_{j'}} \mu_{j,k,j',k'} F^{(j,k)} F^{(j',k')} + \varepsilon.$$

Autant dire qu'on le manipule plutôt numériquement que mathématiquement.

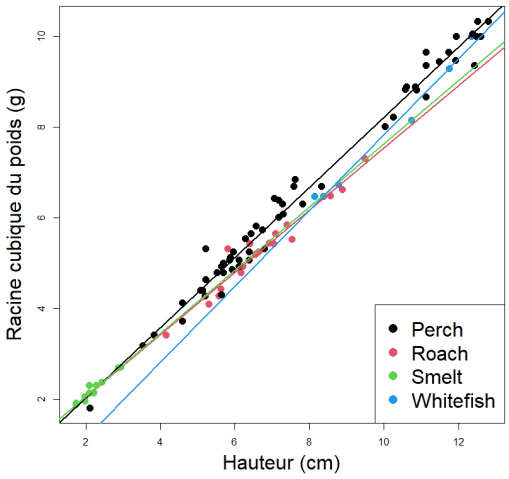
Exemple: On s'intéresse aux données de poids et de hauteurs de 159 poissons, de 7 espèces différentes, pêchés dans le lac Längelmävesi en Finlande.
(Source : Brofeldt, Pekka : *Bidrag till kaennedom on fiskbestondet i vaera sjoear. Laengelmaevesi. T.H.Jaervi : Finlands Fiskeriet Band 4, Meddelanden utgivna av fiskerifoereningen i Finland.* **Helsingfors** 1917)



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Height	1	627.3	627.3	6626.60	<2e-16	***
Species	6	259.5	43.3	456.92	<2e-16	***
Height:Species	6	16.1	2.7	28.39	<2e-16	***
Residuals	144	13.6	0.1			



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Height	1	123.37	123.37	1551.319	<2e-16	***
Species	1	0.03	0.03	0.334	0.566	
Height:Species	1	0.14	0.14	1.799	0.187	
Residuals	42	3.34	0.08			



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Height	1	537.6	537.6	8748.885	< 2e-16	***
Species	3	2.7	0.9	14.732	7.73e-08	***
Height:Species	3	0.3	0.1	1.616	0.191	
Residuals	87	5.3	0.1			

Chapitre 2

Modèles linéaires généralisés

Introduction

Pour les modèles linéaires on a vu que les y_i sont considérés comme réalisations de variable indépendantes

$$Y_i \sim \mathcal{N}\left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j}, \sigma^2\right).$$

On généralise ce principe en considérant maintenant que les Y_i sont indépendants de loi $\mathbb{P}$ appartenant à une famille de lois $\mathcal{F}$ et d'espérance $f(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j})$ pour une fonction f inversible. On a donc

$$\theta_0 + \sum_{j=1}^{d} \theta_j x_{i,j} = f^{-1}(\mathbb{E}[Y_i]).$$

La fonction f^{-1} est alors appelée le **lien** car elle relie la combinaison linéaire des variables explicatives avec l'espérance de la variable d'intérêt.

Un modèle généralisé est donc défini par une famille de lois $\mathcal{F}$ et une fonction de lien f.

Exemples:

- Le modèle linéaire multiple que l'on a étudié dans la première partie du cours est un modèle Gaussiens avec lien identité.
- Le modèle

$$Y_i \sim \mathcal{P}\left(e^{\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j}}\right) \Rightarrow \theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} = \log(\mathbb{E}[Y_i])$$

  est appelé modèle de Poisson avec lien logarithme.

Remarque: On a toujours besoin de l'hypothèse que X_e est de rang plein afin d'avoir l'identifiabilité du modèle.

I Modèle logistique

1 Présentation du modèle

On considère que Y est une variable discrète à deux modalités que l'on représente par 0 et 1. La loi des Y_i est donc une loi de Bernoulli dont la fonction de lien f^{-1} est une fonction inversible de $[0,1] \to \mathbb{R}$. Le choix classique est la fonction **logit** :

$$f^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

dont l'inverse est la fonction sigmoïde (sigmoïde = qui a une forme de S) :

$$f(x) = \frac{e^x}{1+e^x}.$$

On a donc $Y_i \sim b(\pi_i(\theta))$ où

$$\pi_i(\theta) = \mathbb{E}[Y_i] = f\left(\theta_0 + \sum_{j=1}^{d} \theta_j x_{i,j}\right) = \frac{e^{\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j}}}{1 + e^{\theta_0 + \sum_{j=1}^{d} \theta_j x_{i,j}}} = \frac{e^{(X_e\theta)_i}}{1+e^{(X_e\theta)_i}}.$$

On notera $\pi(\theta) \in \mathbb{R}^n$ le vecteur des $\pi_i(\theta)$ par la suite. On a donc un modèle de Bernoulli avec lien logit appelé le **modèle logistique**. On peut interpréter le choix de la fonction de lien en utilisant la définition suivante.

Définition 48

*On appelle la **cote** d'un individu i la quantité*

$$\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)} = \frac{\pi_i(\theta)}{1-\pi_i(\theta)} = e^{(X_e\theta)_i}.$$

Remarques:

- Si $(X_e\theta)_i > 0$ alors $\mathbb{P}(Y_i = 1) > \mathbb{P}(Y_i = 0)$.
- Si $(X_e\theta)_i < 0$ alors $\mathbb{P}(Y_i = 1) < \mathbb{P}(Y_i = 0)$.
- Si $(X_e\theta)_i = 0$ alors $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = 0) = \frac{1}{2}$.
- On en déduit aussi que si on augmente la valeur de la j-ème variable de l'individu i par une constante C alors sa cote va être multipliée par $e^{\theta_j C}$. Cela signifie donc qu'un effet additif sur les variables explicatives va avoir un effet multiplicatif sur la cote des individus. Il en vient l'interprétation suivante des θ_i :
  - Si $\theta_i = 0$ alors la variable explicative associée au coefficient n'a aucun effet sur la variable d'intérêt.
  - Si $\theta_i > 0$ alors une augmentation de la variable explicative associée au coefficient va entraîner une augmentation de $\mathbb{P}(Y_i = 1)$ (et inversement).
  - Si $\theta_i < 0$ alors une augmentation de la variable explicative associée au coefficient va entraîner une diminution de $\mathbb{P}(Y_i = 1)$ (et inversement).

2 Estimation des paramètres par maximum de vraisemblance

Les Y_i sont indépendants et vérifient

$$\mathbb{P}(Y_i = y_i) = \pi_i(\theta)^{y_i}(1-\pi_i(\theta))^{1-y_i} = \begin{cases} \pi_i(\theta) & \text{si } y_i = 1; \\ 1-\pi_i(\theta) & \text{si } y_i = 0. \end{cases}$$

La vraisemblance du modèle est donc

$$L(\theta|Y) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi_i(\theta)^{y_i}(1-\pi_i(\theta))^{1-y_i}$$

et la log-vraisemblance est

$$\begin{aligned}\mathcal{L}(\theta|Y) &= \sum_{i=1}^{n} \left(y_i \log(\pi_i(\theta)) + (1-y_i)\log(1-\pi_i(\theta))\right) \\ &= \sum_{i=1}^{n} \left(y_i \log\left(\frac{\pi_i(\theta)}{1-\pi_i(\theta)}\right) + \log(1-\pi_i(\theta))\right) \\ &= \sum_{i=1}^{n} \left(y_i (X_e\theta)_i - \log\left(1+e^{(X_e\theta)_i}\right)\right).\end{aligned}$$

On remarque que

$$\begin{cases} \dfrac{\partial (X_e\theta)_i}{\partial \theta_0} = 1; \\ \dfrac{\partial (X_e\theta)_i}{\partial \theta_j} = x_{i,j} \quad \text{si} \quad j \neq 0; \end{cases}$$

Par abus de notation on va noter $x_{i,0} = 1$ pour tout i d'où

$$\frac{\partial (X_e\theta)_i}{\partial \theta_j} = x_{i,j} \text{ pour tout } j \text{ et } X_e = \begin{pmatrix} x_{1,0} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,0} & \cdots & x_{n,p} \end{pmatrix} = (x_{i,j})_{\substack{1\leqslant i\leqslant n \\ 0\leqslant j\leqslant p}}$$

On en déduit alors

$$\frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_j} = \sum_{i=1}^{n} \left(y_i x_{i,j} - \frac{x_{i,j} e^{(X_e\theta)_i}}{1+e^{(X_e\theta)_i}}\right) = \sum_{i=1}^{n} \left(x_{i,j}(y_i - \pi_i(\theta))\right) = ({}^tX_e(Y-\pi(\theta)))_j.$$

donc $\nabla\mathcal{L}(\theta|Y) = {}^tX_e(Y-\pi(\theta))$ et

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(\theta|Y)}{\partial\theta_j\partial\theta_k} &= \sum_{i=1}^{n} - \left(\frac{x_{i,j}x_{i,k}e^{(X_e\theta)_i}(1+e^{(X_e\theta)_i}) - x_{i,j}e^{(X_e\theta)_i}x_{i,k}e^{(X_e\theta)_i}}{(1+e^{(X_e\theta)_i})^2}\right) \\ &= \sum_{i=1}^{n} - \left(\frac{x_{i,j}x_{i,k}e^{(X_e\theta)_i}}{(1+e^{(X_e\theta)_i})^2}\right) \\ &= -({}^tX_eD(\theta)X_e)_{j,k} \quad \text{où} \quad D(\theta) = \begin{pmatrix} \frac{e^{(X_e\theta)_1}}{(1+e^{(X_e\theta)_1})^2} & & 0 \\ & \ddots & \\ 0 & & \frac{e^{(X_e\theta)_n}}{(1+e^{(X_e\theta)_n})^2} \end{pmatrix}.\end{aligned}$$

On obtient $H(\mathcal{L})(\theta|Y) = -{}^tX_eD(\theta)X_e$ qui est bien une matrice définie négative quel que soit θ car $D(\theta)$ est définie positive (car diagonale à coefficients > 0) et X_e est de rang plein par hypothèse. Si cette matrice est inversible pour tout θ alors la log-vraisemblance est une fonction strictement concave et admet donc un unique maximum au point qui annule sa dérivée. L'EMV $\hat{\theta}$ vérifie donc l'équation

$${}^tX_e(Y-\pi(\hat{\theta})) = 0_{p+1}$$

appelée **équation de score**.

Cette équation n'est pas résolvable mise à part dans des cas particuliers. On doit donc utiliser une méthode numérique pour obtenir une approximation de la solution. La méthode classique est la méthode de Newton basée sur le développement de Taylor d'ordre 1 de $\nabla\mathcal{L}(\theta|Y)$:

$$\nabla\mathcal{L}(\hat{\theta}|Y) \approx \nabla\mathcal{L}(\theta|Y) + H(\mathcal{L})(\theta|Y)(\hat{\theta}-\theta) \;\Rightarrow\; \hat{\theta} \approx \theta - H(\mathcal{L})(\theta|Y)^{-1}\nabla\mathcal{L}(\theta|Y) \text{ car } \nabla\mathcal{L}(\hat{\theta}|Y) = 0.$$

$\hat{\theta}$ est donc limite de la suite $(\theta_n)_{n\in\mathbb{N}}$ définie par $\theta_{n+1} = \theta_n - H(\mathcal{L})(\theta_n|Y)^{-1}\nabla\mathcal{L}(\theta_n|Y)$. En pratique on se donne donc une valeur initiale θ_0 et on calcule les valeurs θ_n jusqu'à ce que $\|\theta_n - \theta_{n-1}\| \leqslant \varepsilon$ pour un niveau de précision $\varepsilon > 0$ fixé à l'avance. On approxime alors $\hat{\theta}$ par θ_n.

3 Loi des paramètres, intervalles de confiance et tests

Théorème 49

Sous de "bonnes conditions" sur un modèle statistique de paramètre $\theta \in \mathbb{R}^d$ et de n données, l'EMV $\hat{\theta}$ vérifie

$$I_n(\hat{\theta})^{1/2}\left(\hat{\theta}-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0_d, I_d)$$

où $I_n(\theta)$ est appelée la ***matrice d'information de Fisher*** *définie par*

$$I_n(\theta) = \mathbb{E}[-H(\mathcal{L})(\theta|Y)].$$

Autrement dit, si le nombre de données est assez grandes alors on peut approximer la loi de $\hat{\theta}$ par une loi $\mathcal{N}_d(\theta, I_n(\hat{\theta})^{-1})$. Dans le cas du modèle logistique on va donc approximer la loi de $\hat{\theta}$ par un loi $\mathcal{N}_{p+1}(\theta, \Sigma)$ où $\Sigma = ({}^tX_eD(\hat{\theta})X_e)^{-1}$.
Pour tout i on a donc l'approximation

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{\Sigma_{i+1,i+1}}} \sim \mathcal{N}(0,1).$$

On en déduit donc un intervalle de confiance asymptotique de niveau $1-\alpha$ pour les θ_i :

$$IC_{1-\alpha}(\theta_i) = \left[\hat{\theta}_i \pm z_{1-\alpha/2}\sqrt{\Sigma_{i+1,i+1}}\right],$$

où z_α le quantile d'ordre α de la loi $\mathcal{N}(0,1)$. On en déduit aussi que l'on rejette l'hypothèse $\mathcal{H}_0 : \{\theta_i = 0\}$ au risque α si la statistique

$$Z = \frac{\hat{\theta}_i}{\sqrt{\Sigma_{i+1,i+1}}}$$

vérifie $|Z| \geqslant z_{1-\alpha/2}$.

Maintenant, on cherche à tester la nullité d'une ou plusieurs combinaisons linéaires des coefficients, c'est à dire $\mathcal{H}_0 : \{C\theta = 0_k\}$ où C est une matrice de taille $k \times (p+1)$ de rang plein. Comme on peut approximer la loi de $\hat{\theta}$ par une loi $\mathcal{N}_{p+1}(\theta, \Sigma)$ alors, sous $\mathcal{H}_0$, on peut approximer la loi de $C\hat{\theta}$ par une loi $\mathcal{N}_k(0_k, C\Sigma{}^tC)$ et la loi de la statistique $S = \|(C\Sigma{}^tC)^{-1/2}C\hat{\theta}\|^2$ par une loi $\chi^2(k)$. De plus, S va prendre de plus grandes valeurs sous $\mathcal{H}_1$. On va donc rejeter $\mathcal{H}_0$ au risque α si S est plus grand que le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k)$. C'est ce qu'on appelle le **test de Wald**.

4 Cas particuliers

- On appelle **modèle vide** le modèle ne contenant aucune variable et juste l'intercept. Dans ce cas-là, $X_e = 1_n$, $\theta = \theta_0$, $\pi(\theta) = \frac{e^{\theta_0}}{1-e^{\theta_0}}1_n$ et l'équation de score devient

$$\begin{aligned}
{}^t1_n\left(Y - \frac{e^{\hat{\theta}_0}}{1+e^{\hat{\theta}_0}}1_n\right) = 0 &\Leftrightarrow \sum_{i=1}^n\left(y_i - \frac{e^{\hat{\theta}_0}}{1+e^{\hat{\theta}_0}}\right) = 0 \\
&\Leftrightarrow \sum_{i=1}^n y_i = n\frac{e^{\hat{\theta}_0}}{1+e^{\hat{\theta}_0}} \\
&\Leftrightarrow \frac{e^{\hat{\theta}_0}}{1+e^{\hat{\theta}_0}} = \bar{Y}.
\end{aligned}$$

Comme l'inverse de la fonction sigmoïde $f(x) = \frac{e^x}{1+e^x}$ est la fonction logit $f^{-1}(x) = \log(x/(1-x))$ alors on en déduit

$$\hat{\theta}_0 = \log\left(\frac{\bar{Y}}{1-\bar{Y}}\right).$$

La log-vraisemblance du modèle en $\hat{\theta}_0$ s'écrit alors

$$\begin{aligned}\mathcal{L}_{vide}(\hat{\theta}_0|Y) &= \sum_{i=1}^{n}\Big(y_i\log(\pi_i(\hat{\theta})) + (1-y_i)\log(1-\pi_i(\hat{\theta}))\Big)\\ &= \sum_{i=1}^{n}\Big(y_i\log(\bar{Y}) + (1-y_i)\log(1-\bar{Y})\Big)\\ &= n\Big(\bar{Y}\log(\bar{Y}) + (1-\bar{Y})\log(1-\bar{Y})\Big).\end{aligned}$$

- On appelle **modèle saturé** le modèle contenant $p = n-1$ variables quantitatives. C'est le nombre maximum possible de variables que l'on peut avoir de sorte que les colonnes de X_e forment une famille libre. Dans ce cas-là, la matrice X_e est de taille $n \times n$ et inversible. L'équation de score devient alors

$$^tX_e(Y - \pi(\hat{\theta}_{\text{sat}})) = 0_{p+1} \;\Leftrightarrow\; Y = \pi(\hat{\theta}_{\text{sat}}).$$

Cela ne permet pas d'obtenir la valeur des $\hat{\theta}_{\text{sat}}$ mais on arrive quand même à obtenir la relation $\pi_i(\hat{\theta}_{\text{sat}}) = y_i$ pour tout i. En particulier, la vraisemblance en $\hat{\theta}_{\text{sat}}$ pour le modèle saturé s'écrit

$$\begin{aligned}\mathcal{L}_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y) &= \sum_{i=1}^{n}\Big(y_i\log(\pi_i(\hat{\theta}_{\text{sat}})) + (1-y_i)\log(1-\pi_i(\hat{\theta}_{\text{sat}}))\Big)\\ &= \sum_{i=1}^{n}\left(y_i\log(y_i) + (1-y_i)\log(1-y_i)\right)\\ &= 0.\end{aligned}$$

A la dernière partie on a utilisé le fait que $y_i = 0$ ou 1 et $1\log(1) = 0 = \lim_{x\to 0} x\log(x)$.

5 Qualité d'ajustement et sélection de modèle

Comme le modèle saturé donne le meilleur ajustement possible et donc la vraisemblance la plus faible possible alors il est naturel de juger la qualité d'ajustement d'un modèle en utilisant l'écart entre sa log-vraisemblance et celle du modèle saturé. De même, le modèle vide (et donc sa vraisemblance) correspond au pire cas possible. Pour un modèle donnée on veut donc que la valeur de sa vraisemblance en l'EMV soit la plus proche de celle du modèle saturé et la plus éloignée de celle du modèle vide. Il en vient la définition suivante.

Définition 50

On définit la ***déviance*** *d'un modèle généralisé par la quantité*

$$D = -2(\log L(\hat{\theta}|Y) - \log L_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y)) = -2(\mathcal{L}(\hat{\theta}|Y) - \mathcal{L}_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y)) \geqslant 0.$$

On définit la ***déviance nulle*** *d'un modèle généralisé par la quantité*

$$D_0 = -2(\log L_{\text{vide}}(\hat{\theta}_0|Y) - \log L_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y)) = -2(\mathcal{L}_{\text{vide}}(\hat{\theta}_0|Y) - \mathcal{L}_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y)) \geqslant 0.$$

On a $0 \leqslant D \leqslant D_0$. Plus la déviance est proche de 0, plus le modèle est bien ajusté. A l'inverse, plus la déviance est proche de D_0 plus le modèle est similaire au modèle vide.

Qualité du modèle

Une approximation (un peu douteuse mathématiquement) est que si le modèle est bien ajusté alors la loi de D doit être proche d'une loi $\chi^2(n-(p+1))$. On va donc tester l'hypothèse qu'on a un bon ajustement en comparant D aux quantiles de la loi $\chi^2(n-(p+1))$. De façon similaire, la loi de $D_0 - D$ est en général approximé par une loi $\chi^2(p)$ et on teste alors s'il y a une différence significative entre notre modèle et le modèle vide en comparant $D_0 - D$ aux quantiles de la loi $\chi^2(p)$. Comme la moyenne d'une loi $\chi^2(n)$ est n alors un critère alternatif et plus vague est de dire qu'on a un mauvais ajustement si

$$\frac{D}{n-(p+1)} \gg 1$$

et on a un modèle significativement différent du modèle vide si

$$\frac{D_0 - D}{p} \gg 1.$$

Comparaison de modèles

On considère une famille de modèles $\mathcal{M}$ et pour tout modèle m on note sa déviance $D(m)$. Si on considère deux modèles m_1 et m_2 tels que $m_1 \subset m_2$ alors on considère que le modèle m_1 est significativement meilleur que le modèle m_2 si

$$\frac{D(m_1) - D(m_2)}{|m_2| - |m_1|} \gg 1.$$

On peut aussi utiliser un test statistique en approximant la loi de $D(m_1) - D(m_2)$ par une loi $\chi^2(|m_2| - |m_1|)$.

Pour faire de la sélection de modèle, on peut toujours utiliser les méthodes numériques (validation croisée et non-croisées) vues précédemment ou les critères AIC et BIC :

$$AIC = -2\log L(\hat{\theta}|Y) + 2(p+1) \quad \text{et} \quad BIC = -2\log L(\hat{\theta}|Y) + (p+1)\log(n).$$

Quitte à rajouter le terme constant $2\log L_{\text{sat}}(\hat{\theta}_{\text{sat}})$ on peut aussi les écrire

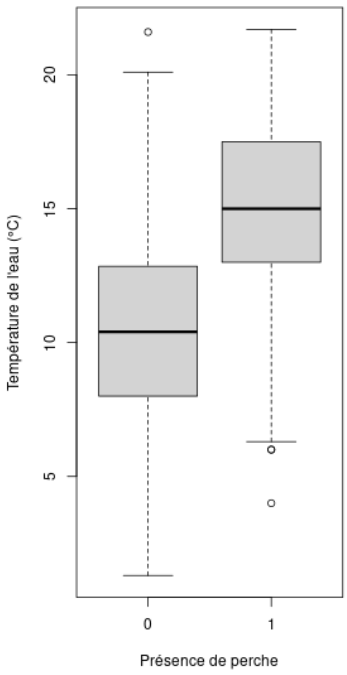
$$AIC(m) = D(m) + 2(p+1) \quad \text{et} \quad BIC(m) = D(m) + (p+1)\log(n).$$

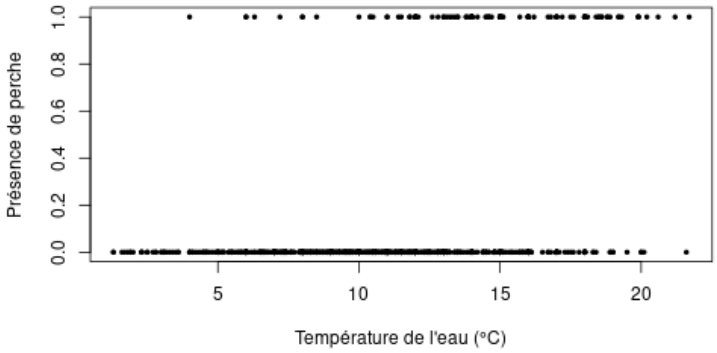
6 Exemple d'application sur R

On s'intéresse aux données de l'article suivant :
(Sutela, T. and Vehanen, T. and Jounela, P. and Aroviita, J. (2021). *Species–environment relationships of fish and map-based variables in small boreal streams : Linkages with climate change and bioassessment*. Ecology and Evolution, **11**, 10457–10467)
Ce sont 776 données de présence/absence de différent types de poissons dans des petit cours d'eau en Finlande accompagné de variables environnementales sur ces cours d'eaux. On commencer par modéliser la présence de Perche (1 pour présence et 0 pour absence) en fonction de la température (en °C) du cours d'eau.

Water.temperature.at.sampling	Perch
0	8.40
0	12.50
0	12.00
0	15.80
1	15.10
0	10.30
⋮	⋮

On peut représenter visuellement ces données avec des boîtes à moustache ou un nuage de points.

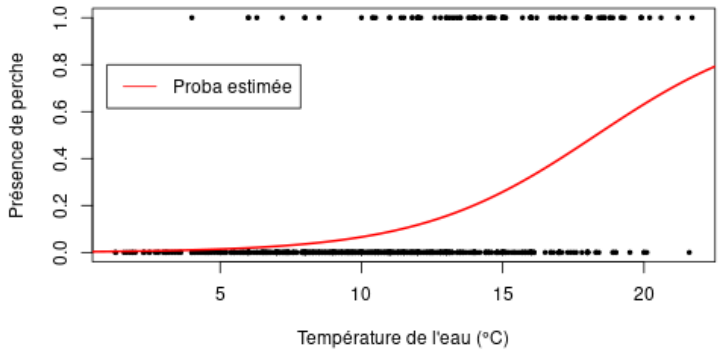




Visuellement, on observe que plus la température de l'eau augmente plus les chances d'observer une perche augmente. On ajuste alors un modèle logistique de la forme

$$\text{Présence de perche} \sim b(\pi(\theta_0 + \theta_1 \times \text{Température})) \text{ avec } \pi(x) = \frac{e^x}{1+e^x}.$$

On obtient alors $\hat{\theta}_0 \approx -5.84$ et $\hat{\theta}_1 \approx 0.32$. Le fait que $\hat{\theta}_1$ est positif confirme qu'une augmentation de la température de l'eau augmente les chances d'observer une perche. On peut alors visualiser la probabilité d'observation d'une perche estimée par le modèle en superposant le graphe de la fonction $x \mapsto \hat{\pi}(x) = \pi(\hat{\theta}_0 + \hat{\theta}_1 x)$ au nuage de point des données.



On s'intéresse maintenant à modéliser la présence de Perches en fonction de plusieurs variables : l'aire de la zone de capture (en km^2), la latitude, la température de l'eau (en °C), l'altitude (en m), la proportion de zone urbaine autour du cours d'eau (en %) et la proportion de lac autour du cours d'eau (en %).

Perch	Catchment.area	Latitude	Water.temperature.at.sampling	Altitude	Urban.areas	Lakes
0	1.09	7253482	8.40	150.26	0.00	0.00
0	1.09	7253482	12.50	150.26	0.00	0.00
0	1.14	7254200	12.00	234.50	0.00	3.06
0	1.15	6679943	15.80	4.13	71.24	0.00
1	1.15	6679907	15.10	3.37	71.24	0.00
0	1.19	7297771	10.30	242.67	0.00	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Le résultat de l'ajustement d'un modèle logistique sur R est donné çi-dessous.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3016  -0.4648  -0.2851  -0.1533   3.1228

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 6.043e+00  6.568e+00   0.920  0.35758
Catchment.area              1.739e-02  6.090e-03   2.856  0.00429 **
Latitude                   -1.638e-06  9.885e-07  -1.657  0.09759 .
Water.temperature.at.sampling 2.750e-01  3.576e-02   7.690 1.47e-14 ***
Altitude                   -4.861e-03  3.169e-03  -1.534  0.12507
Urban.areas                -2.574e-02  9.444e-03  -2.726  0.00641 **
Lakes                       1.097e-01  2.683e-02   4.089 4.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 629.81  on 775  degrees of freedom
Residual deviance: 459.77  on 769  degrees of freedom
AIC: 473.77

Number of Fisher Scoring iterations: 6
```

→ Ecart-type des $\hat{\theta}_i$
→ p-valeur du test de significativité du modèle
→ Statistique Z du test de nullité des $\hat{\theta}_i$
→ Estimation des θ_i.
→ D_0
→ $n-1$
→ $n-(p+1)$
→ D
→ AIC

On observe les résultats suivant :

- On a $459.77 = D < n-(p+1) = 769$ et $170.04 = D_0 - D \gg p = 6$ donc le modèle est bien ajusté et il est significativement meilleur que le modèle vide.
- Les variables ayant un effet significatif sur la présence de Perche sont l'aire de la zone de capture, la température de l'eau et la proportion de lacs et de zones urbaines autour de la zone de capture.
- Au vu des signes des coefficients de ces variables on en déduit que, aux autres effets fixés, les chances d'observer une perche augmentent lorsque la zone de capture est élevée, la

température est élevée, il y a beaucoup de lacs autour de la zone de capture mais peu de zones urbaines.

7 Le rapport des cotes

a Le cas univarié à 2 modalités

De façon similaire à un modèle ANOVA, on considère qu'on a une seule variable explicative X qualitative à deux modalités A et B. On note $\mathbb{1}_{i,A}$ et $\mathbb{1}_{i,B}$ les variables qui valent 1 si l'individu i possède la modalité A ou B de la variable explicative et 0 sinon. Les données peuvent alors être représentées par une table de contingence.

Y/X	A	B	Total
0	$n_{0,A}$	$n_{0,B}$	n_0
1	$n_{1,A}$	$n_{1,B}$	n_1
Total	n_A	n_B	n

De façon similaire à un ANOVA, on considère alors le modèle sans intercept

$$Y_i \sim b(\pi_i(\theta)) \text{ où } \pi_i(\theta) = \frac{e^{\theta_A \mathbb{1}_{i,A} + \theta_B \mathbb{1}_{i,B}}}{1 + e^{\theta_A \mathbb{1}_{i,A} + \theta_B \mathbb{1}_{i,B}}}.$$

Pour un individu i possédant la modalité A on a donc $\pi_i(\theta) = \frac{e^{\theta_A}}{1+e^{\theta_A}}$ et sa cote est

$$\frac{\mathbb{P}(Y_i = 1 | x_i = A)}{\mathbb{P}(Y_i = 0 | x_i = A)} = e^{\theta_A}.$$

S'il possède la modalité B on a donc $\pi_i(\theta) = \frac{e^{\theta_B}}{1+e^{\theta_B}}$ et sa cote est e^{θ_B}.

Définition 51

On appelle rapport des cotes (Odds Ratio) entre la modalité A et B de X la quantité

$$\text{OR} = \frac{\frac{\mathbb{P}(Y_i=1|x_i=A)}{\mathbb{P}(Y_i=0|x_i=A)}}{\frac{\mathbb{P}(Y_i=1|x_i=B)}{\mathbb{P}(Y_i=0|x_i=B)}} = \frac{e^{\theta_A}}{e^{\theta_B}} = e^{\theta_A - \theta_B}.$$

Remarque: Le rapport des cotes permet de modéliser l'effet de la variable explicative sur la variable d'intérêt :

- Si OR $= 1$ alors la variable explicative n'a aucune influence sur la cote de Y.
- Si OR > 1 alors la cote de Y est plus élevée pour les individus avec la modalité A que pour les individus avec la modalité B.
- Si OR < 1 alors la cote de Y est plus élevée pour les individus avec la modalité B que pour les individus avec la modalité A.

Pour ce modèle, on a plusieurs simplifications dans le calcul de la vraisemblance qui amène le résultat suivant.

Proposition 52

L'estimateur du maximum de vraisemblance de $\hat{\theta}_A$ et $\hat{\theta}_B$ vérifie

$$e^{\hat{\theta}_A} = \frac{n_{1,A}}{n_{0,A}} \text{ et } e^{\hat{\theta}_B} = \frac{n_{1,B}}{n_{0,B}}.$$

Démonstration : Si on reprend l'équation de score sous sa forme non vectorielle on a

$$\sum_{i=1}^{n}\Big(x_{i,j}(y_i-\pi_i(\hat{\theta}))\Big)=0.$$

Or, ici soit $x_{i,j}$ correspond à $\mathbb{1}_{i,A}$ et dans ce cas-là on a $\pi_i(\hat{\theta})=\frac{e^{\hat{\theta}_A}}{1+e^{\hat{\theta}_A}}$ soit $x_{i,j}$ correspond à $\mathbb{1}_{i,B}$ et dans ce cas-là on a $\pi_i(\hat{\theta})=\frac{e^{\hat{\theta}_B}}{1+e^{\hat{\theta}_B}}$. Dans le premier cas, l'équation de score devient

$$\sum_{i=1}^{n}\left(\mathbb{1}_{i,A}(y_i-\frac{e^{\hat{\theta}_A}}{1+e^{\hat{\theta}_A}}\right)=0\ \Leftrightarrow\ \sum_{i:\mathbb{1}_{i,A}=1}\left(y_i-\frac{e^{\hat{\theta}_A}}{1+e^{\hat{\theta}_A}}\right)=0\ \Leftrightarrow\ n_{1,A}-n_A\frac{e^{\hat{\theta}_A}}{1+e^{\hat{\theta}_A}}=0.$$

On en déduit alors

$$\frac{e^{\hat{\theta}_A}}{1+e^{\hat{\theta}_A}}=\frac{n_{1,A}}{n_A}\ \Leftrightarrow\ \hat{\theta}_A=\text{logit}\left(\frac{n_{1,A}}{n_A}\right)=\log\left(\frac{\frac{n_{1,A}}{n_A}}{1-\frac{n_{1,A}}{n_A}}\right)=\log\left(\frac{\frac{n_{1,A}}{n_A}}{\frac{n_{0,A}}{n_A}}\right)=\log\left(\frac{n_{1,A}}{n_{0,A}}\right).$$

Avec le même raisonnement on va aussi obtenir

$$\frac{e^{\hat{\theta}_B}}{1+e^{\hat{\theta}_B}}=\frac{n_{1,B}}{n_B}\ \Leftrightarrow\ \hat{\theta}_B=\log\left(\frac{n_{1,B}}{n_{0,B}}\right).$$

■

Ce résultat nous donne directement l'estimateur des cotes des individus possédant la modalité A et des individus possédant la modalité B et il en découle l'estimateur suivant pour le rapport des cotes :

$$\widehat{\text{OR}}=e^{\hat{\theta}_A-\hat{\theta}_B}=\frac{\frac{n_{1,A}}{n_{0,A}}}{\frac{n_{1,B}}{n_{0,B}}}=\frac{n_{1,A}n_{0,B}}{n_{0,A}n_{1,B}}.$$

On cherche maintenant à tester l'hypothèse que la modalité A ou B de la variable explicative n'influence pas la valeur de la variable d'intérêt, c'est à dire :

$$\mathcal{H}_0=\{\theta_A=\theta_B\}\text{ contre }\mathcal{H}_1=\{\hat{\theta}_A\neq\hat{\theta}_B\}\quad\Leftrightarrow\quad\mathcal{H}_0=\{\text{OR}=1\}\text{ contre }\mathcal{H}_1=\{\text{OR}\neq1\}.$$

Pour faire ce test on doit d'abord trouver une bonne approximation de la loi de $\hat{\theta}_A-\hat{\theta}_B$. On va alors pouvoir s'aider du fait que la Hessienne de ce modèle possède une expression simplifiée.

Proposition 53

Pour ce modèle on a

$$H(\mathcal{L})(\theta_A,\theta_B|Y)=-\begin{pmatrix}\frac{n_Ae^{\theta_A}}{(1+e^{\theta_A})^2} & 0\\ 0 & \frac{n_Be^{\theta_B}}{(1+e^{\theta_B})^2}\end{pmatrix}.$$

En conséquence,

$$H(\mathcal{L})(\hat{\theta}_A,\hat{\theta}_B|Y)=-\begin{pmatrix}\frac{n_{0,A}n_{1,A}}{n_A} & 0\\ 0 & \frac{n_{0,B}n_{1,B}}{n_B}\end{pmatrix}.$$

Démonstration : En reprenant la formule de la Hessienne sous sa forme non vectorielle on a

$$\frac{\partial^2\mathcal{L}(\theta|Y)}{\partial\theta_A\partial\theta_A}=\sum_{i=1}^{n}-\left(\frac{\mathbb{1}_{i,A}\mathbb{1}_{i,A}e^{(X_e\theta)_i}}{(1+e^{(X_e\theta)_i})^2}\right)=-\frac{n_Ae^{\theta_A}}{(1+e^{\theta_A})^2}.$$

De même,

$$\frac{\partial^2\mathcal{L}(\theta|Y)}{\partial\theta_B\partial\theta_B}=-\frac{n_Be^{\theta_B}}{(1+e^{\theta_B})^2}\quad\text{et}\quad\frac{\partial^2\mathcal{L}(\theta|Y)}{\partial\theta_A\partial\theta_B}=\sum_{i=1}^{n}-\left(\frac{\mathbb{1}_{i,A}\mathbb{1}_{i,B}e^{(X_e\theta)_i}}{(1+e^{(X_e\theta)_i})^2}\right)=0$$

car $\mathbb{1}_{i,A}\mathbb{1}_{i,B} = 0$ pour tout i. De plus, comme $e^{\hat{\theta}_A} = \frac{n_{1,A}}{n_{0,A}}$ alors

$$\frac{n_A e^{\hat{\theta}_A}}{(1+e^{\hat{\theta}_A})^2} = \frac{n_A \frac{n_{1,A}}{n_{0,A}}}{(1+\frac{n_{1,A}}{n_{0,A}})^2} = \frac{n_A \frac{n_{1,A}}{n_{0,A}}}{(\frac{n_A}{n_{0,A}})^2} = \frac{n_{0,A} n_{1,A}}{n_A}$$

et avec le même raisonnement on obtient $\frac{n_B e^{\hat{\theta}_B}}{(1+e^{\hat{\theta}_B})^2} = \frac{n_{0,B} n_{1,B}}{n_B}$. ■

On en déduit alors qu'on peut approximer la loi de $(\hat{\theta}_A, \hat{\theta}_B)$ par une loi

$$\mathcal{N}_2\left(\begin{pmatrix}\theta_A\\ \theta_B\end{pmatrix}, \begin{pmatrix}\frac{n_A}{n_{0,A}n_{1,A}} & 0\\ 0 & \frac{n_B}{n_{0,B}n_{1,B}}\end{pmatrix}\right)$$

et donc

$$\begin{aligned}\hat{\theta}_A - \hat{\theta}_B &\sim \mathcal{N}\left(\theta_A - \theta_B, \frac{n_A}{n_{0,A}n_{1,A}} + \frac{n_B}{n_{0,B}n_{1,B}}\right)\\ &\sim \mathcal{N}\left(\theta_A - \theta_B, \frac{1}{n_{0,A}} + \frac{1}{n_{1,A}} + \frac{1}{n_{0,B}} + \frac{1}{n_{1,B}}\right).\end{aligned}$$

Comme $\hat{\theta}_A - \hat{\theta}_B = \log(\widehat{\mathrm{OR}})$ alors la statistique

$$Z = \left(\frac{1}{n_{0,A}} + \frac{1}{n_{1,A}} + \frac{1}{n_{0,B}} + \frac{1}{n_{1,B}}\right)^{-1/2} \log(\widehat{\mathrm{OR}})$$

a une loi $\mathcal{N}(0,1)$ sous l'hypothèse $\mathcal{H}_0$ que $\hat{\theta}_A = \hat{\theta}_B$. On rejette donc l'hypothèse $\mathcal{H}_0$ au risque α lorsque $|Z| \geqslant z_{1-\alpha/2}$. De façon similaire on peut aussi obtenir l'intervalle de confiance suivant pour OR :

$$\left[e^{\log(\widehat{\mathrm{OR}}) - z_{1-\alpha/2}\sqrt{\frac{1}{n_{0,A}}+\frac{1}{n_{1,A}}+\frac{1}{n_{0,B}}+\frac{1}{n_{1,B}}}}, e^{\log(\widehat{\mathrm{OR}}) + z_{1-\alpha/2}\sqrt{\frac{1}{n_{0,A}}+\frac{1}{n_{1,A}}+\frac{1}{n_{0,B}}+\frac{1}{n_{1,B}}}}\right]$$

Exemple: On s'intéresse à l'effet d'un traitement d'une maladie comparé à un placebo. On rapport çi-dessous les résultats de tests sur 100 patients.

	Traitement	Placebo
Guéri	30	20
Non Guéri	10	40

Si on considère la guérison comme le cas où $Y = 1$ alors la cote "Guéri" contre "Non Guéri" de la modalité *Traitement* est estimée par $\frac{30}{10} = 3$ et la cote de la modalité *Placebo* est estimée par $\frac{20}{40} = 0.5$ et le rapport des cotes est alors

$$\widehat{\mathrm{OR}} = \frac{\frac{30}{10}}{\frac{20}{40}} = 6.$$

Un intervalle de confiance pour le rapport des cotes à 95% est donc

$$\left[e^{\log(6) \pm 1.96\sqrt{\frac{1}{10}+\frac{1}{20}+\frac{1}{30}+\frac{1}{40}}}\right] \approx [2.45, 14.68].$$

La statistique de test est alors

$$Z = \left(\frac{1}{10} + \frac{1}{20} + \frac{1}{30} + \frac{1}{40}\right)^{-1/2} \log(6) \approx 3.93 > 1.96$$

donc on rejette au risque 5% l'hypothèse que le traitement n'a pas d'effet sur la guérison.

Remarque: Comparé au test du χ^2 qui cherche une dépendance générale entre deux variables, le rapport des cotes permet de quantifier plus précisément comment les modalités d'une variables affecte celles d'une autre. Par contre, les rapports des cotes ne concerne que des variables avec deux modalités.

b Le cas univarié avec au moins 3 modalités

On considère maintenant que la variable explicative a trois modalités A, B et C. Le modèle s'écrit alors

$$Y_i \sim b(\pi_i(\theta)) \text{ où } \pi_i(\theta) = \frac{e^{\theta_A \mathbb{1}_{i,A} + \theta_B \mathbb{1}_{i,B} + \theta_C \mathbb{1}_{i,C}}}{1 + e^{\theta_A \mathbb{1}_{i,A} + \theta_B \mathbb{1}_{i,B} + \theta_C \mathbb{1}_{i,C}}}$$

Comme $\mathbb{1}_{i,A} + \mathbb{1}_{i,B} + \mathbb{1}_{i,C} = 1$ alors

$$\theta_A \mathbb{1}_{i,A} + \theta_B \mathbb{1}_{i,B} + \theta_C \mathbb{1}_{i,C} = \theta_A + (\theta_B - \theta_A)\mathbb{1}_{i,B} + (\theta_C - \theta_A)\mathbb{1}_{i,C}$$

En posant $\mu_0 = \theta_A$, $\mu_1 = (\theta_B - \theta_A)$ et $\mu_2 = (\theta_C - \theta_A)$ on se ramène alors à un modèle logistique avec un intercept :

$$Y_i \sim b(\pi_i(\mu)) \text{ où } \pi_i(\theta) = \frac{e^{\mu_0 + \mu_1 \mathbb{1}_{i,B} + \mu_2 \mathbb{1}_{i,C}}}{1 + e^{\mu_0 + \mu_1 \mathbb{1}_{i,B} + \mu_2 \mathbb{1}_{i,C}}}.$$

De plus, on a directement que e^{μ_1} correspond au rapport des cotes entre les modalités B et A et e^{μ_2} correspond au rapport des cotes entre les modalités C et A. La modalité A est alors appelée la modalité de référence. On peut aussi tester si $\theta_A = \theta_B = \theta_C$ comme pour un ANOVA en testant si $\mu_1 = \mu_2 = 0$ avec un test de Wald.

c Le cas multivarié

On considère le cas général où on a un nombre quelconque de variables explicatives qualitatives et quantitative. Pour chaque variable qualitative, on met de côté une modalité (la modalité de référence) et on ajoute au modèle les indicatrices des autres modalités. L'exponentielle de ces coefficients est alors appelé le **rapport des cotes ajusté** de ces modalités par rapport à la modalité de référence. Pour calculer les intervalles de confiance et faire les tests on remplace alors dans les formules précédentes les quantités $\frac{1}{n_{0,A}} + \frac{1}{n_{1,A}} + \frac{1}{n_{0,B}} + \frac{1}{n_{1,B}}$ par les écarts-types estimés des coefficients.

8 Extensions du modèle logistique

⚠ On a vu que toutes les propriétés des modèles généralisés peuvent se déduire principalement de l'équation de score et de la Hessienne de la log-vraisemblance. On se contentera donc de calculer ces quantités (quand c'est possible) pour les modèles que l'on verra à partir de maintenant.

Le modèle binomial

Une première généralisation du modèle logistique est le modèle binomial avec lien logit qui suppose que les y_i sont issus de variables aléatoires

$$Y_i \sim B(N_i, \pi_i(\theta)) \quad \text{où} \quad \pi_i(\theta) = \frac{e^{\theta_0 + \sum_{j=1}^d \theta_j X_{i,j}}}{1 + e^{\theta_0 + \sum_{j=1}^d \theta_j X_{i,j}}} = \frac{e^{(X_e\theta)_i}}{1 + e^{(X_e\theta)_i}}$$

et les N_i sont des valeurs connues.

Proposition 54

- *L'équation de score du modèle est*

$$ {}^tX_e\Big(Y - D(N)\pi(\theta)\Big) = 0_{p+1} $$

  où $D(N)$ est la matrice diagonale des N_i.
- *La Hessienne de la log-vraisemblance du modèle s'écrit*

$$ H(\mathcal{L})(\theta|Y) = -{}^tX_eD(N)D(\theta)X_e \implies I_n(\theta) = {}^tX_eD(N)D(\theta)X_e. $$

Démonstration : La vraisemblance du modèle est

$$ L(\theta|Y) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \binom{N_i}{y_i} \pi_i(\theta)^{y_i}(1-\pi_i(\theta))^{N_i - y_i} $$

et la log-vraisemblance est

$$ \begin{aligned} \mathcal{L}(\theta|Y) &= \sum_{i=1}^{n} \left(\log\left(\binom{N_i}{y_i}\right) + y_i \log(\pi_i(\theta)) + (N_i - y_i)\log(1-\pi_i(\theta)) \right) \\ &= \sum_{i=1}^{n} \left(\log\left(\binom{N_i}{y_i}\right) + y_i \log\left(\frac{\pi_i(\theta)}{1-\pi_i(\theta)}\right) + N_i \log(1-\pi_i(\theta)) \right) \\ &= \sum_{i=1}^{n} \left(\log\left(\binom{N_i}{y_i}\right) + y_i (X_e\theta)_i - N_i \log\left(1 + e^{(X_e\theta)_i}\right) \right). \end{aligned} $$

En reprenant la convention $x_{i,0} = 1$ pour tout i on obtient

$$ \frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_j} = \sum_{i=1}^{n} \left(y_i x_{i,j} - N_i \frac{x_{i,j} e^{(X_e\theta)_i}}{1 + e^{(X_e\theta)_i}} \right) = \sum_{i=1}^{n} (x_{i,j}(y_i - N_i\pi_i(\theta))) = ({}^tX_e(Y - D(N)\pi(\theta)))_j, $$

où $D(N)$ est la matrice diagonale des N_i. On a donc $\nabla\mathcal{L}(\theta) = {}^tX_e(Y - D(N)\pi(\theta))$ et on peut montrer que $H(\mathcal{L})(\theta) = -{}^tX_eD(N)D(\theta)X_e$ en reprenant le même raisonnement que pour le modèle logistique. ■

Remarque: Si on remplace l'individu i par N_i individus avec les mêmes valeurs pour les variables explicatives et tel que y_i individus ont la valeur 1 pour la variable d'intérêt et $N_i - y_i$ ont la valeur 0 alors on se ramène à un modèle logistique qui est équivalent au modèle binomial.

Le modèle multinomial

On considère que Y est une variables discrète avec M valeurs possibles que l'on notera $1, \cdots, M$. On note $\pi_i^{(j)} = \mathbb{P}(Y_i = j)$ et on souhaite modéliser chaque $\pi_i^{(j)}$ comme une fonction d'une combinaison linéaire des variables explicatives sous la contrainte évidente que $\pi_i^{(1)} + \cdots + \pi_i^{(M)} = 1$. Le modèle est alors décrit par $M-1$ probabilités. On va donc paramétriser les probabilités $\pi_i^{(1)}, \cdots, \pi_i^{(M-1)}$ et poser $\pi_i^{(M)} = 1 - \sum_{j=1}^{M-1} \pi_i^{(j)}$. A chaque probabilité $\pi_i^{(j)}$ pour $j \in \{1, \cdots, M-1\}$ va être associé un ensemble de paramètres $\theta_0^{(j)}, \cdots, \theta_p^{(j)}$. On note alors $\theta^{(j)}$ le vecteur des $\theta_i^{(j)}$. Ce modèle possède donc $(M-1)(p+1)$ paramètres.

Dans le modèle logistique on modélisait le rapport $\frac{\mathbb{P}(Y_i=1)}{\mathbb{P}(Y_i=0)}$ (la côte) par $e^{(X_e\theta)_i}$ ce qui conduisait au lien logit. De façon similaire, dans le modèle multinomial on fait la modélisation suivante :

$$ \forall j \in \{1, \cdots, M-1\}, \ \frac{\mathbb{P}(Y_i = j)}{\mathbb{P}(Y_i = M)} = \frac{\pi_i^{(j)}}{\pi_i^{(M)}} = e^{(X_e\theta^{(j)})_i}. $$

On obtient donc :

$$\forall j \in \{1, \cdots, M-1\},\ \pi_i^{(j)} = \frac{e^{(X_e\theta^{(j)})_i}}{1 + \sum_{k=1}^{M-1} e^{(X_e\theta^{(k)})_i}} \text{ et } \pi_i^{(M)} = \frac{1}{1 + \sum_{k=1}^{M-1} e^{(X_e\theta^{(k)})_i}}.$$

On retrouve bien le modèle logistique pour $M = 2$.

Proposition 55

Les dérivées partielles du modèle multinomial s'écrivent

$$\frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_j^{(k)}} = \sum_{i=1}^{n} x_{i,j} \left(\mathbb{1}_{y_i=k} - \pi_i^{(k)} \right)$$

où on pose $x_{i,0} = 0$ pour tout i.

Démonstration : La vraisemblance du modèle est

$$L(\theta|Y) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi_i^{(y_i)}.$$

Si on pose $\theta^{(M)} = 0_{p+1}$ alors on peut écrire $\pi_i^{(y_i)} = e^{(X_e\theta^{(y_i)})_i}\pi_i^{(M)}$ d'où

$$\begin{aligned}
\mathcal{L}(\theta|Y) = \sum_{i=1}^{n} \log(\pi_i^{(y_i)}) &= \sum_{i=1}^{n} \left(\log(e^{(X_e\theta^{(y_i)})_i}) + \log(\pi_i^{(M)}) \right) \\
&= \sum_{i=1}^{n} \left((X_e\theta^{(y_i)})_i - \log\left(1 + \sum_{k=1}^{M-1} e^{(X_e\theta^{(k)})_i}\right)\right) \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{M-1} (X_e\theta^{(k)})_i \mathbb{1}_{y_i=k} - \log\left(1 + \sum_{k=1}^{M-1} e^{(X_e\theta^{(k)})_i}\right)\right).
\end{aligned}$$

Les dérivées partielles de $\mathcal{L}(\theta|Y)$ s'expriment alors par

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_j^{(k)}} &= \sum_{i=1}^{n} \left(x_{i,j}\mathbb{1}_{y_i=k} - \frac{x_{i,j}e^{(X_e\theta^{(k)})_i}}{1 + \sum_{l=1}^{M-1} e^{(X_e\theta^{(l)})_i}} \right) \\
&= \sum_{i=1}^{n} x_{i,j} \left(\mathbb{1}_{y_i=k} - \pi_i^{(k)} \right)
\end{aligned}$$

ce qui nous donne l'équation de score. ■

II Le modèle de Poisson et ses variantes

1 Le modèle de Poisson classique

On considère que Y est une variable à valeur dans $\mathbb{N}$ que l'on va modéliser par une loi de Poisson. La fonction de lien est une fonction inversible de $\mathbb{R}_+^*$ dans $\mathbb{R}$. Le choix classique est la fonction **logarithme**, c'est à dire que $(X_e\theta)_i = \log(\mathbb{E}[Y_i])$. On a donc

$$Y_i \sim \mathcal{P}(\lambda_i(\theta)) \quad \text{où} \quad \lambda_i(\theta) = e^{(X_e\theta)_i} = e^{\theta_0 + \sum_{j=1}^{d} \theta_j X_{i,j}}.$$

On a donc un modèle de Poisson avec lien logarithmique. On notera $\lambda(\theta)$ pour le vecteur des $\lambda_i(\theta)$.

Remarque: Si on augmente la valeur de la j-ème variable de l'individu i par une constante C alors sa moyenne va être multipliée par $e^{\theta_j C}$. Cela signifie donc qu'un effet additif sur les variables explicatives va avoir un effet multiplicatif sur la valeur moyenne des Y_i.

Proposition 56

- *L'équation de score du modèle est*

$$^tX_e(Y - \lambda(\theta)) = 0_{p+1}.$$

- *La Hessienne de la log-vraisemblance du modèle est*

$$-(^tX_eD(\theta)X_e)_{j,k} \quad où \quad D(\theta) = \begin{pmatrix} e^{(X_e\theta)_1} & & 0 \\ & \ddots & \\ 0 & & e^{(X_e\theta)_n} \end{pmatrix}.$$

Démonstration : La vraisemblance du modèle est

$$L(\theta|Y) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \frac{\lambda_i(\theta)^{y_i} e^{-\lambda_i(\theta)}}{y_i!}$$

et la log-vraisemblance est

$$\mathcal{L}(\theta|Y) = \sum_{i=1}^{n} (y_i \log(\lambda_i(\theta)) - \lambda_i(\theta) - \log(y_i!)) = \sum_{i=1}^{n} \left(y_i (X_e\theta)_i - e^{(X_e\theta)_i} - \log(y_i!) \right).$$

En réutilisant le fait que $\frac{\partial (X_e\theta)_i}{\partial \theta_j} = x_{i,j}$, où on pose $x_{i,0} = 1$ pour tout i, on obtient

$$\frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_j} = \sum_{i=1}^{n} \left(y_i x_{i,j} - x_{i,j} e^{(X_e\theta)_i} \right) = \sum_{i=1}^{n} (x_{i,j}(y_i - \lambda_i(\theta))) = (^tX_e(Y - \lambda(\theta)))_j.$$

donc $\nabla_\theta \mathcal{L}(\theta) = {}^tX_e(Y - \lambda(\theta))$ et

$$\frac{\partial^2 \mathcal{L}(\theta|Y)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^{n} -x_{i,j} x_{i,k} e^{(X_e\theta)_i} = -(^tX_eD(\theta)X_e)_{j,k}$$

■

Remarque: Comme $D(\theta)$ est tout le temps définie positive et X_e est de rang plein alors $-{}^tX_eD(\theta)X_e$ est donc tout le temps définie négative. La log-vraisemblance est donc strictement concave. De plus, on en déduit que $I_n(\theta) = {}^tX_eD(\theta)X_e$ donc on approximera la loi de $\hat{\theta}$ par une loi $\mathcal{N}_{p+1}(\theta, (^tX_eD(\theta)X_e)^{-1})$.

A partir de ce résultat on peut en déduire les propriétés du modèle de Poisson de façon similaire à ce qu'on a fait pour le modèle logistique. En particulier, on a les résultats suivants permettant le calcul des déviances :

Proposition 57

- *La log-vraisemblance du modèle vide en l'EMV est :*

$$\mathcal{L}_{\text{vide}}(\hat{\theta}_0|Y) = n\bar{Y}\log(\bar{Y}) - n\bar{Y} - \sum_{i=1}^{n} \log(y_i!).$$

- *La log-vraisemblance du modèle saturé en l'EMV est :*

$$\mathcal{L}_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y) = \sum_{i=1}^{n} y_i \log(y_i) - n\bar{Y} - \sum_{i=1}^{n} \log(y_i!).$$

Démonstration : • Pour le modèle vide on a $\lambda_i(\theta) = e^{\theta_0}$ et $X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ d'où

$$\frac{\partial \mathcal{L}(\theta|Y)}{\partial \theta_0} = \sum_{i=1}^{n} \left(y_i - e^{\theta_0} \right) = n(\bar{Y} - e^{\theta_0}).$$

L'EMV est donc $\hat{\theta}_0 = \log(\bar{Y})$ d'où

$$\mathcal{L}_{\text{vide}}(\hat{\theta}_0|Y) = \sum_{i=1}^{n} \left(y_i\hat{\theta}_0 - e^{\hat{\theta}_0} - \log(y_i!) \right)$$
$$= n\bar{Y}\hat{\theta}_0 - ne^{\hat{\theta}_0} - \sum_{i=1}^{n} \log(y_i!) = n\bar{Y}\log(\bar{Y}) - n\bar{Y} - \sum_{i=1}^{n} \log(y_i!)$$

• Pour le modèle saturé sans répétition on a que X_e est inversible. Comme $\nabla_\theta \mathcal{L}(\theta) = {}^tX_e(Y - \lambda(\theta))$ on en déduit alors que $\lambda(\hat{\theta}_{\text{sat}}) = Y$ et donc $\lambda_i(\hat{\theta}_{\text{sat}}) = y_i$ pour tout i d'où

$$\mathcal{L}_{\text{sat}}(\hat{\theta}_{\text{sat}}|Y) = \sum_{i=1}^{n} \left(y_i \log(\lambda_i(\hat{\theta}_{\text{sat}})) - \lambda_i(\hat{\theta}_{\text{sat}}) - \log(y_i!) \right) = \sum_{i=1}^{n} y_i \log(y_i) - n\bar{Y} - \sum_{i=1}^{n} \log(y_i!).$$

■

2 Le modèle de Poisson avec inflation de zéros (ZIP)

Pour une loi de Poisson on a égalité entre l'espérance et la variance ce qui est une forte contrainte. On va donc s'intéresser à deux modèles permettant de modifier la variance d'une loi de Poisson sans changer son espérance. Tout d'abord, si Y a tendance à souvent prendre la valeur 0 on peut modifier le modèle Poisson en ce qu'on appelle un modèle de Poisson avec inflation de zéro.

Définition 58

*On dit qu'une variable aléatoire Y suit une loi de **Poisson avec inflation de zéros** de paramètres $p \in [0,1]$ et $\lambda \in \mathbb{R}_+^*$, notée $ZIP(p,\lambda)$, si elle prend la valeur 0 avec probabilité p et suit une loi de Poisson $\mathcal{P}(\lambda)$ avec probabilité $1-p$ (indépendante du choix). On a alors*

$$\begin{cases} \mathbb{P}(Y=0) = p + (1-p)e^{-\lambda}; \\ \mathbb{P}(Y=k) = (1-p)\frac{\lambda^k e^{-\lambda}}{k!} \text{ pour } k \in \mathbb{N}\backslash\{0\}. \end{cases}$$

Proposition 59

Si $Y \sim ZIP(p,\lambda)$ alors

$$\mathbb{E}[Y] = \lambda(1-p) \quad et \quad Var(Y) = \lambda(1-p)(1+p\lambda)$$

Démonstration : Soient $B \sim b(p)$ et Y égal à 0 conditionnellement à $B = 1$ et de loi $\mathcal{P}(\lambda)$ indépendante à B conditionnellement à $B = 0$. Par la formule de l'espérance total on a

$$\mathbb{E}[X] = \mathbb{E}[X|B=0]\mathbb{P}(X=0) + \mathbb{E}[X|B=1]\mathbb{P}(X=1) = \lambda(1-p) + 0$$

et

$$\mathbb{E}[X^2] = \mathbb{E}[X^2|B=0]\mathbb{P}(X=0) + \mathbb{E}[X^2|B=1]\mathbb{P}(X=1) = (\lambda+\lambda^2)(1-p) + 0$$

d'où

$$\text{Var}(X) = (\lambda+\lambda^2)(1-p) - \lambda^2(1-p)^2 = \lambda(1-p)(1+\lambda-\lambda(1-p)) = \lambda(1-p)(1+p\lambda).$$

■

On considère maintenant deux jeux de données X et Z de variables explicatives (certaines pouvant être en commun). On peut alors modéliser la loi des Y_i par une loi $ZIP(\pi_i(\theta), \lambda_i(\mu))$ où

$$\begin{cases} \lambda_i(\mu) = e^{(X_e\mu)_i} & \text{lien } \textbf{log}; \\ \pi_i(\theta) = \frac{e^{(Z_e\theta)_i}}{1+e^{(Z_e\theta)_i}} & \text{lien } \textbf{logit}. \end{cases}$$

Les paramètres peuvent alors être estimés par maximum de vraisemblance.

3 La loi quasi-Poisson

Le principe du modèle quasi-Poisson est de modifier la vraisemblance d'une loi de Poisson de sorte à changer sa variance sans changer son espérance. On a vu que la log-vraisemblance d'une observation y d'une loi de Poisson $\mathcal{P}(\lambda)$ est

$$\mathcal{L}(\lambda|y) = y\log(\lambda) - \lambda - \log(y!).$$

On considère maintenant la vraisemblance modifiée suivante :

$$\mathcal{L}(\lambda, \phi|y) = \frac{y\log(\lambda) - \lambda}{\phi} + f(\phi, y),$$

où $\phi > 0$ est un nouveau paramètre, f est une fonction de $\mathbb{R}_+ \times \mathbb{N}$ dans $\mathbb{R}$ choisie de sorte que l'on aie bien la log-vraisemblance d'une loi de probabilité pour tout λ et ϕ, c'est à dire

$$\sum_{n=0}^{+\infty} \exp\left(\frac{n\log(\lambda) - \lambda}{\phi} + f(\phi, n)\right) = 1.$$

On admet qu'une telle fonction existe et on appelle loi **Quasi-Poisson** la loi sur $\mathbb{N}$ de probabilités

$$\mathbb{P}(X = n) = \exp\left(\frac{n\log(\lambda) - \lambda}{\phi} + f(\phi, n)\right).$$

On note $X \sim Q\mathcal{P}(\lambda, \phi)$.

Proposition 60

Si $X \sim Q\mathcal{P}(\lambda, \phi)$ alors

$$\mathbb{E}[X] = \lambda \text{ et } Var(X) = \phi\lambda.$$

Démonstration : Si on dérive l'expression de la somme des probabilités étant égale à 1 par rapport à λ alors on obtient

$$\sum_{n=0}^{+\infty} \left(\frac{n}{\lambda\phi} - \frac{1}{\phi}\right) \exp\left(\frac{n\log(\lambda) - \lambda}{\phi} + f(\phi, n)\right) = 0 \Leftrightarrow \frac{\mathbb{E}[X]}{\lambda\phi} - \frac{1}{\phi} = 0$$

d'où $\mathbb{E}[X] = \lambda$. Si on redérive l'expression précédente par rapport à λ alors on obtient.

$$\sum_{n=0}^{+\infty} -\frac{n}{\lambda^2\phi} \exp\left(\frac{n\log(\lambda) - \lambda}{\phi} + f(\phi, n)\right) + \sum_{n=0}^{+\infty} \left(\frac{n}{\lambda\phi} - \frac{1}{\phi}\right)^2 \exp\left(\frac{n\log(\lambda) - \lambda}{\phi} + f(\phi, n)\right) = 0$$

$$\Leftrightarrow -\frac{1}{\lambda^2\phi}\mathbb{E}[X] + \frac{1}{\lambda^2\phi^2}\mathbb{E}[(X-\lambda)^2] = 0$$

$$\Leftrightarrow -\phi\lambda + \mathrm{Var}(X) = 0$$

d'où $\mathrm{Var}(X) = \phi\lambda = \phi\mathbb{E}[X]$. ■

La loi Quasi-Poisson possède donc deux paramètres permettant de contrôler à la fois son espérance et sa variance. On considère maintenant que l'on a p variables explicatives quantitatives et on suppose que les y_i sont issues de variables aléatoires $Y_i \sim \mathcal{P}(\lambda_i(\theta), \phi)$ où $\lambda_i(\theta) = e^{(X_e\theta)_i}$. Dans ce cas-là, le paramètre θ s'estime de la même façon que pour le modèle de Poisson.

Proposition 61

- *L'équation de score du modèle est*

$$ {}^tX_e(Y - \lambda(\theta)) = 0_{p+1}. $$

- *La Hessienne de la log-vraisemblance du modèle est*

$$ -\frac{1}{\phi}({}^tX_e D(\theta) X_e)_{j,k} \quad \text{où} \quad D(\theta) = \begin{pmatrix} e^{(X_e\theta)_1} & & 0 \\ & \ddots & \\ 0 & & e^{(X_e\theta)_n} \end{pmatrix}. $$

Démonstration : Comme

$$ \mathcal{L}(\theta, \phi | Y) = \sum_{i=1}^{n} \left(\frac{y_i \log(\lambda_i(\theta)) - \lambda_i(\theta)}{\phi} + f(\phi, y_i) \right) $$

alors le gradient par rapport à θ est

$$ \nabla_\theta \mathcal{L}(\theta, \phi | Y) = \nabla_\theta \left(\sum_{i=1}^{n} \left(\frac{y_i \log(\lambda_i(\theta)) - \lambda_i(\theta)}{\phi} \right) \right) = \frac{1}{\phi} {}^tX_e(Y - \lambda(\theta)) $$

et de Hessienne par rapport à θ

$$ H_\theta(\mathcal{L})(\theta, \phi | Y) = -\frac{1}{\phi} {}^tX_e D(\theta) X_e. $$

L'équation de score du modèle s'écrit donc

$$ {}^tX_e(Y - \lambda(\theta)) = 0_{p+1} $$

et c'est exactement la même que pour le modèle de Poisson. ■

Par contre, on ne va pas pouvoir estimer ϕ par maximum de vraisemblance. On défini alors la statistique

$$ S^2 = \sum_{i=1}^{n} \frac{(Y_i - \lambda_i(\hat{\theta}))^2}{\lambda_i(\hat{\theta})}. $$

Comme $\mathbb{E}[Y_i] = \lambda_i(\theta)$ et $\mathrm{Var}(Y_i) = \phi\lambda_i(\theta)$ alors il est courant (mais un peu douteux) d'approximer la loi de S^2/ϕ par une loi $\chi^2(n - (p+1))$. On estime alors ϕ par

$$ \hat{\phi} = \frac{S^2}{n - (p+1)}. $$

On remarque que la matrice d'information de Fisher pour θ est

$$ I_n(\theta, \phi) = -\mathbb{E}[H_\theta(\mathcal{L})(\theta, \phi | Y)] = \frac{1}{\phi} {}^tX_e D(\theta) X_e. $$

Du coup, si on note $\Sigma = (X_e D(\hat{\theta}) X_e)^{-1}$ alors on va pouvoir approximer asymptotiquement la loi de $\hat{\theta}$ par une loi $\mathcal{N}_{p+1}(\theta, \phi\Sigma)$. Avec cette approximation et en supposant que la loi de $\hat{\theta}$ et $\hat{\phi}$ est indépendante (ce qui est pas vrai mais on approxime encore) on a

$$ \frac{\hat{\theta}_i - \theta_i}{\sqrt{\phi\Sigma_{i+1,i+1}}} \sim \mathcal{N}(0,1) \Rightarrow \frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{\phi}\Sigma_{i+1,i+1}}} \sim \mathcal{T}(n - (p+1)). $$

C'est le résultat que l'on utilise pour faire des intervalles de confiances et des tests de nullité.

Enfin, on considère la déviance du modèle sans remplacer ϕ par son estimateur :

$$D = -2(\mathcal{L}(\hat{\theta}, \phi|Y) - \mathcal{L}_{sat}(\hat{\theta}_{sat}, \phi|Y)).$$

Comme cette quantité est proportionnelle à $\frac{1}{\phi}$ alors on note ϕD la "scaled deviance" qui est donc une quantité calculable et vérifie

$$\phi D = -2(\mathcal{L}(\hat{\theta}, 1|Y) - \mathcal{L}_{sat}(\hat{\theta}_{sat}, 1|Y))$$

ce qui correspond à la déviance pour le modèle de Poisson. En approximant la loi de D par une loi $\chi^2(n-(p+1))$, la loi de $D_0 - D$ par une loi $\chi^2(p)$ et la loi de $\frac{(n-(p+1))\hat{\phi}}{\phi}$ par une loi $\chi^2(n-(p+1))$ on obtient :

$$\frac{\phi D}{(n-(p+1))\hat{\phi}} \sim \mathcal{F}(n-(p+1), n-(p+1)) \text{ et } \frac{\phi(D_0 - D)}{p\hat{\phi}} \sim \mathcal{F}(p, n-(p+1)).$$

On utilise alors ces quantités pour tester si le modèle est bien ajusté et s'il est significativement meilleur que le modèle nul.

III Introduction au modèle à effets mixtes

Exemple : On considère qu'on observe des quantités $(x_{i,j}, y_{i,j})$ pour un individu i à plusieurs temps t_j. On ne peut pas directement appliquer le modèle linéaire à X et Y vu qu'il va y avoir de la dépendance entre les données issues d'un même individu. Rajouter le numéro de l'individu comme variable qualitative est aussi douteux si on a peu de temps t_j différents car le modèle va se retrouver avec trop de paramètres. Une solution est d'associer à chaque individu un "effet aléatoire" :

$$Y_{i,j} = \theta_0 + x_{i,j}\theta_1 + \alpha_i + \varepsilon_{i,j}$$

où les $\varepsilon_{i,j}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et les α_i sont i.i.d. de loi $\mathcal{N}(0, \tau^2)$ et indépendant des $\varepsilon_{i,j}$. Le terme $\theta_0 + x_{i,j}\theta_1$ est appelé **l'effet fixe** et le terme α_i est appelé **l'effet aléatoire**. On obtient alors

$$\mathrm{Var}(Y_{i,j}) = \mathrm{Var}(\alpha_i) + \mathrm{Var}(\varepsilon_{i,j}) = \sigma^2 + \tau^2$$

et

$$\mathrm{cov}(Y_{i,j}, Y_{i',j'}) = \begin{cases} \sigma^2 + \tau^2 \text{ si } i = i' \text{ et } j = j' \\ \tau^2 \text{ si } i = i' \text{ et } j \neq j' \\ 0 \text{ sinon.} \end{cases}$$

De façon plus général, on considère qu'on a une variable d'intéret Y, des variables explicatives X modélisant les effets fixes et J variables qualitatives modélisant les effets aléatoires. On note $q_1, \cdots, q_J$ leur nombre de modalités et $K = q_1 + \cdots + q_J$. On associe à la k-ième modalité de la j-ème variable un vecteur $Z_i^{(j,k)}$ tel que $Z_i^{(j,k)} = 1$ si l'individu i possède la modalité et 0 sinon. Le modèle mixte s'écrit

$$Y_i = (X_e\theta)_i + \sum_{j=1}^{J}\sum_{k=1}^{q_j} Z_i^{(j,k)}\alpha_{j,k} + \varepsilon_i$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et les $\alpha_{j,k}$ sont indépendant de loi $\mathcal{N}(0, \tau_j^2)$ et indépendant des ε_i. Si on note Z la matrice dont les colonnes sont les $Z^{(j,k)}$ et α le vecteur des $\alpha_{j,k}$ alors ce modèle se réécrit vectoriellement sous la forme

$$Y = X_e\theta + Z\alpha + \varepsilon$$

où $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ et $\alpha \sim \mathcal{N}_K(0_K, D)$ où D est une matrice diagonale contenant q_1 fois la valeur τ_1^2 puis q_2 fois la valeur τ_2^2, etc... Les paramètres du modèle sont donc θ, σ^2 et les τ_i^2. De plus,

$$\mathrm{Var}(Y) = \mathrm{Var}(Z\alpha) + \mathrm{Var}(\varepsilon) = \sigma^2 I_n + ZD^tZ$$

d'où $Y \sim \mathcal{N}_n(X_e\theta, \sigma^2 I_n + ZD^tZ)$. Si on note $V = \sigma^2 I_n + ZD^tZ$ alors on peut obtenir des équations compliquées à résoudre numériquement pour estimer σ^2 et les τ_i^2 par maximum de vraisemblance. Une fois qu'on les a on peut obtenir une estimation $\hat{V}$ de cette matrice de covariance et l'estimateur de θ par maximum de vraisemblance s'écrit

$$\hat{\theta} = ({}^tX\hat{V}^{-1}X)^{-1\,t}X\hat{V}^{-1}Y.$$

IV Les familles implémentées dans la fonction *glm*

On termine par une description rapide des familles et fonctions de liens implémentées dans la fonction *glm* de R.

- **binomial**
  Lien par défaut : logit
  Autres liens disponibles : probit, cauchit, log, cloglog
  On a $Y_i \sim B(N, \pi_i(\theta))$. Soit Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$ et $F(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi}$ la fonction de répartition de la loi de Cauchy de densité $f(x) = \frac{1}{\pi(1+x^2)}$. Φ et F sont bien des fonctions de $\mathbb{R}$ dans $[0,1]$. On a les liens suivant :

$$\begin{cases} \pi_i(\theta) = \frac{e^{(X_e\theta)_i}}{1+e^{(X_e\theta)_i}} & \textbf{logit} \\ \pi_i(\theta) = \Phi((X_e\theta)_i) & \textbf{probit} \\ \pi_i(\theta) = F((X_e\theta)_i) & \textbf{cauchit} \\ \pi_i(\theta) = e^{(X_e\theta)_i} & \textbf{log} \\ \pi_i(\theta) = 1 - e^{-e^{(X_e\theta)_i}} & \textbf{cloglog} \end{cases}$$

  Si on pose $Z_i = (X_e\theta)_i + \varepsilon_i$ où $\varepsilon_i \sim \mathcal{N}(0,1)$ alors le lien probit correspond à avoir $Y_i = \mathbb{1}_{Z_i>0}$ car

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i > 0) = \mathbb{P}(\varepsilon_i > -(X_e\theta)_i) = \mathbb{P}(\varepsilon_i < (X_e\theta)_i) = \Phi((X_e\theta)_i).$$

  Si ε_i suit la loi de Cauchy alors on obtient le modèle cauchit.

- **gaussian**
  Lien par défaut : identity
  Autres liens disponibles : log, inverse
  On a $Y_i \sim \mathcal{N}(\mu_i(\theta), \sigma^2)$ avec les liens suivant :

$$\begin{cases} \mu_i(\theta) = (X_e\theta)_i & \textbf{identity} \\ \mu_i(\theta) = e^{(X_e\theta)_i} & \textbf{log} \\ \mu_i(\theta) = \frac{1}{(X_e\theta)_i} & \textbf{inverse} \end{cases}$$

- **Gamma**
  Lien par défaut : inverse
  Autres liens disponibles : identity, log
  On a $Y_i \sim \Gamma(\lambda_i(\theta), k)$. On rappelle que la loi $\Gamma(\lambda, k)$ est définie pour $\lambda, k > 0$ et a pour densité

$$f(x) = \frac{x^{k-1}e^{-x/\lambda}}{\lambda^k\Gamma(k)}\mathbb{1}_{x>0} \text{ où } \Gamma(k) = \int_0^{+\infty} x^{k-1}e^{-x}\mathrm{d}x.$$

La loi $\Gamma(\lambda, k)$ est d'espérance λk et de variance $\lambda^2 k$. En particulier, si $k = 1$ alors $\Gamma(\lambda, 1)$ est une loi $\mathcal{E}(1/\lambda)$.
On a les liens suivant :

$$\begin{cases} \lambda_i(\theta) = \frac{1}{(X_e\theta)_i} & \textbf{inverse} \\ \lambda_i(\theta) = (X_e\theta)_i & \textbf{identity} \\ \lambda_i(\theta) = e^{(X_e\theta)_i} & \textbf{log} \end{cases}$$

- **inverse.gaussian**
  Lien par défaut : 1/mu^2
  Autres liens disponibles : inverse, identity, log
  On a $Y_i \sim \text{IG}(\mu_i(\theta), \lambda)$ où $\text{IG}(\mu, \lambda)$ est la loi Gaussienne inverse, définie pour $\mu, \lambda > 0$, de densité

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right).$$

  La loi $\text{IG}(\mu, \lambda)$ est d'espérance μ et de variance μ^3/λ. Attention, la loi Gaussienne inverse n'a rien à voir avec l'inverse d'une loi Gaussienne !
  On a les liens suivant :

$$\begin{cases} \mu_i(\theta) = \frac{1}{\sqrt{(X_e\theta)_i}} & \textbf{1/mu\^{}2} \\ \mu_i(\theta) = \frac{1}{(X_e\theta)_i} & \textbf{inverse} \\ \mu_i(\theta) = (X_e\theta)_i & \textbf{identity} \\ \mu_i(\theta) = e^{(X_e\theta)_i} & \textbf{log} \end{cases}$$

- **poisson**
  Lien par défaut : log
  Autres liens disponibles : identity, sqrt
  On a $Y_i \sim \mathcal{P}(\lambda_i(\theta))$ avec les liens suivant :

$$\begin{cases} \lambda_i(\theta) = e^{(X_e\theta)_i} & \textbf{log} \\ \lambda_i(\theta) = (X_e\theta)_i & \textbf{identity} \\ \lambda_i(\theta) = (X_e\theta)_i^2 & \textbf{sqrt} \end{cases}$$

- **quasipoisson**
  Même lien que pour la loi de Poisson mais le paramètre de dispersion ϕ est estimé en plus.
- **quasibinomiale**
  Même lien que pour la loi binomiale mais on rajoute un paramètre de dispersion de façon similaire à ce qu'on a fait pour la loi quasi-Poisson. La log-vraisemblance d'une unique observation y d'une loi $B(N, p)$ est

$$\mathcal{L}(p) = \log\left(\binom{N}{y}\right) + y\log(p) + (N-y)\log(1-p).$$

  Pour le modèle quasi-binomial $QB(N, p, \phi)$ on la transforme en

$$\mathcal{L}(p, \phi) = \frac{y\log(p) + (N-y)\log(1-p)}{\phi} + f(\phi, y).$$

  La loi associée est d'espérance Np et de variance $Np(1-p)\phi$. Pour un modèle $Y_i \sim QB(N, \pi_i(\theta), \phi)$ alors θ est estimé comme pour le modèle binomial et ϕ est estimé par

$$\hat{\phi} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{\left(y_i - N\pi_i(\hat{\theta})\right)^2}{N\pi_i(\hat{\theta})\left(1-\pi_i(\hat{\theta})\right)}$$

- **quasi**
  Lien par défaut : log
  Autres liens disponibles : logit, probit, cloglog, identity, inverse, log, 1/mu^2, sqrt
  Variance par défaut : constant
  Autres variances disponibles : mu(1-mu), mu, mu^2, mu^3

  On considère une loi dont la log-vraisemblance d'une observation y s'écrit

$$\mathcal{L}(\mu,\phi)=\int_y^\mu \frac{y-t}{\phi V(t)}\mathrm{d}t+f(y,\phi) \;\Rightarrow\; \frac{\mathrm{d}\mathcal{L}(\mu,\phi)}{\mathrm{d}\mu}=\frac{y-\mu}{\phi V(t)}.$$

  Alors, sous de faibles hypothèses on peu démontrer que ça définit bien une loi d'espérance μ et de variance $V(\mu)\phi$. Pour un modèle où μ est remplacé par $\mu_i(\theta)$ alors θ est estimé par maximum de vraisemblance et ϕ est estimé par

$$\hat{\phi}=\frac{1}{n-(p+1)}\sum_{i=1}^{n}\frac{(y_i-\mu_i(\hat{\theta}))^2}{V(\mu_i(\hat{\theta}))}$$

  Un tel modèle est donc défini par **une fonction de lien** et **une fonction de variance** V.
  Exemples :
  - Si $V(\mu)=1$ alors

$$\mathcal{L}(\mu,\phi)=\int_y^\mu \frac{y-t}{\phi}\mathrm{d}t+f(y,\phi)=-\frac{(y-\mu)^2}{2\phi}+f(y,\phi)$$

    ce qui corresponds à la log-vraisemblance de la loi $\mathcal{N}(0,1)$ avec $\phi=\sigma^2$:

$$\mathcal{L}(\mu,\phi)=-\frac{1}{2}\log(2\pi\phi)-\frac{(y-\mu)^2}{2\phi}.$$

  - Si $V(\mu)=\mu$ alors

$$\mathcal{L}(\mu,\phi)=\int_y^\mu \frac{y-t}{\phi t}\mathrm{d}t+f(y,\phi)=\frac{y\log(\mu)-\mu}{\phi}+f(y,\phi)$$

    ce qui corresponds à la log-vraisemblance de la loi quasi-Poisson.
  - Si $V(\mu)=N\mu(1-\mu)$ alors

$$\mathcal{L}(\mu,\phi)=\int_y^\mu \frac{y-t}{\phi N t(1-t)}\mathrm{d}t+f(y,\phi)=\frac{y\log(\mu)+(1-y)\log(1-\mu)}{N\phi}+f(y,\phi)$$

    ce qui corresponds à la log-vraisemblance de Y/N pour Y qui suit une loi quasi-binomiale.
  - Si $V(\mu)=\mu^2$ alors

$$\mathcal{L}(\mu,\phi)=\int_y^\mu \frac{y-t}{\phi t^2}\mathrm{d}t+f(y,\phi)=\frac{-\frac{y}{\mu}+\log(\frac{y}{\mu})}{\phi}+\frac{1}{\phi}+f(y,\phi)$$

    On peut vérifier que ça correspond à la log-vraisemblance d'une loi $\Gamma(\lambda,k)$ qui est égale à

$$\mathcal{L}(\lambda,k)=(k-1)\log(y)-\frac{y}{\lambda}-k\log(\lambda)-\log(y)-\log(\Gamma(k))=k\log\left(\frac{y}{\lambda}\right)-\frac{y}{\lambda}-\log(y\Gamma(k)).$$

    En posant $\phi=\frac{1}{k}$ et $\mu=\lambda k=\frac{\lambda}{\phi}$ on obtient bien

$$\mathcal{L}(\mu,\phi)=\frac{-\frac{y}{\mu}+\log(\frac{y}{\mu})}{\phi}-\log(y\Gamma(1/\phi)).$$

— Si $V(\mu) = \mu^3$ alors

$$\mathcal{L}(\mu, \phi) = \int_y^\mu \frac{y-t}{\phi t^3} \mathrm{d}t + f(y, \phi) = \frac{-\frac{y}{2\mu^2} + \frac{1}{\mu}}{\phi} - \frac{1}{2y\phi} + f(y, \phi)$$

On peut vérifier que ça correspond à la log-vraisemblance d'une loi Gaussienne inverse $\mathrm{IG}(\mu, \lambda)$ qui est égale à

$$\mathcal{L}(\mu, \lambda) = -\frac{\lambda(y-\mu)^2}{2\mu^2 y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) = -\frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right).$$

En posant $\phi = \frac{1}{\lambda}$ on obtient bien

$$\mathcal{L}(\mu, \phi) = \frac{-\frac{y}{2\mu^2} + \frac{1}{\mu}}{\phi} - \frac{1}{2y\phi} - \frac{1}{2}\log(2\phi\pi y^3).$$