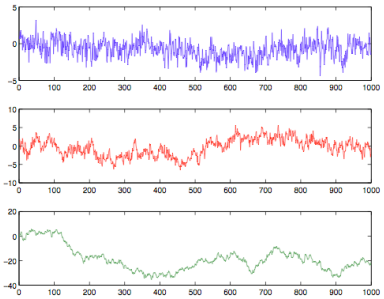
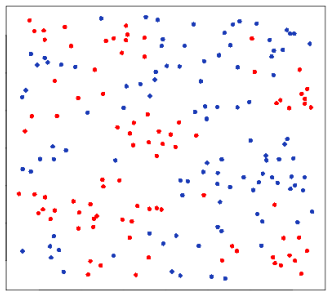
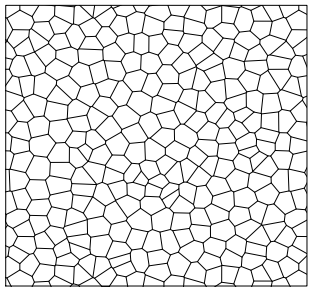
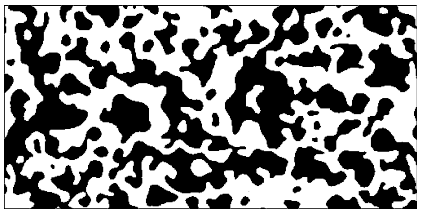
Régression linéaire et généralisée

Frédéric Lavancier

ENSAI 2A
2025-2026

Qui suis-je ?

- Je suis arrivé comme enseignant-chercheur à l'Ensai en 2023.
- Avant j'étais enseignant-chercheur à l'université de Nantes.
- Encore avant, j'étais étudiant à l'Ensai.
- Mon activité de recherche concerne la statistique des processus et la statistique spatiale

- Outre mes activités d'enseignant-chercheur, je suis responsable de la coopération internationale du Genes (Ensai+Ensae)

Intervenants

Cours : Frédéric Lavancier
Emmanuel Pilliat (en anglais)

TD/TP : Julien Jamme
Théo Leroy
Denis Mottin
Marie Christiane Wambo / Koffi Amezouwui
Théo Paquier (en anglais)

Modalités envisagées

- 8 séances de cours
- Un poly et des slides (les deux en évolution) disponibles sur Moodle.
- 8 séances de TD/TP
- Un petit QCM en début de chaque séance de TD/TP
- Un contrôle intermédiaire d'1 heure le 4 novembre (date à confirmer)
- Un examen le 19 décembre (date à confirmer)

Attention : certaines séances de TP pourront avoir lieu en salle de TD avec votre ordinateur personnel (pas la première).

Evaluation

- L'ensemble des QCM formera une note
- Le contrôle intermédiaire formera une autre note
- note de CC=0.5 QCM + 0.5 contrôle
- (les coefficients 0.5-0.5 pourront être ajustés)

$$\boxed{\text{Note finale} = 0.5\ \text{CC} + 0.5\ \text{Examen}}$$

(les coefficients 0.5-0.5 pourront être ajustés)

Plan du cours

Objectifs d'un modèle de régression :

Expliquer une grandeur Y en fonction de p grandeurs $X^{(1)}, \dots, X^{(p)}$ (variables explicatives, ou régresseurs). Pour cela on dispose de n observations de chaque grandeur auprès de n individus.

Exemples :

- Y : la consommation électrique quotidienne en France
  $X = X^{(1)}$: température moyenne journalière.
  Les données sont un historique de Y et X sur n jours.
  Question : a-t-on $Y \approx f(X)$ pour une certaine fonction f ?
  En simplifiant : a-t-on $Y \approx aX + b$ pour certaines valeurs a et b ?
  Si oui, $a =?$, $b =?$ La relation est-elle "fiable" ?

- $Y = 0$ *ou* 1 : qualité d'un client (1 : bon ; 0 : pas bon)
  $X^{(1)}$: revenu du client
  $X^{(2)}$: catégorie socio professionnelle (6-7 possibilités)
  $X^{(3)}$: âge
  Données : n clients.
  On modélise dans ce cas $p = \mathbb{P}(Y = 1)$.
  A-t-on $p \approx f(X^{(1)}, X^{(2)}, X^{(3)})$ pour une fonction f à valeurs dans $[0, 1]$?

La relation "approximative" que l'on cherche à établir entre Y et $X^{(1)}, \ldots, X^{(p)}$ est un **modèle**.

Pourquoi chercher à établir un tel modèle ? Deux raisons principales :

- Objectif descriptif : quantifier l'effet marginal de chaque variable. Par exemple, si $X^{(1)}$ augmente de 10%, comment évolue Y ?
- Objectif prédictif : étant données des nouvelles valeurs pour $X^{(1)}, \ldots, X^{(p)}$, on peut en déduire le Y (approximatif) associé.

Plan du cours

1. Introduction
   $\rightarrow$ Analyse bivariée (rappels) : lien entre 2 variables
   $\rightarrow$ Aspects généraux sur la modélisation

2. Régression linéaire
   $\rightarrow$ Y quantitative en fonction de $X^{(1)}, \ldots, X^{(p)}$ quantitatives

3. Analyse de la variance et de la covariance
   $\rightarrow$ Y quanti en fonction de $X^{(1)}, \ldots, X^{(p)}$ qualitatives et/ou quantitatives

4. Modèles linéaires généralisés
   $\rightarrow$ Y quali ou quanti en fonction de $X^{(1)}, \ldots, X^{(p)}$ quali et/ou quanti

Références

- "Régression avec R", P-A. Cornillon, E. Matzner-Løber
  $\longrightarrow$ *Livre en français, très accessible, en lien avec les 3 premiers chapitres*

- "Le modèle linéaire par l'exemple", J.-M. Azais, J.-M. Bardet.
  $\longrightarrow$ *Livre en français, en lien avec les 3 premiers chapitres : des discussions intéressantes sur l'enjeu des hypothèses, et des résultats théoriques fins.*

- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
  $\longrightarrow$ *Grand classique sur les méthodes de machine learning, y compris les méthodes vues dans ce cours. Exemples avec R.*

- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.
  $\longrightarrow$ *Grand classique également. Version plus théorique (et plus complète) que le précédent.*

Références spécifiques au chapitre 4

- Agresti, A. Foundations of Linear and Generalized Linear Models, Wiley.
  $\longrightarrow$ *Livre classique sur le sujet, en lien avec le chapitre 4*

- Antoniadis, A. Berruyer J. et Carmona R. Régression non linéaire et applications, Economica.
  $\longrightarrow$ *Résultats théoriques complets, en lien avec le chapitre 4*

- Dobson, A.J., Barnett, A.G. An Introduction to Generalized Linear Models, CRC Press.
  $\longrightarrow$ *Des exemples en R, en lien avec le chapitre 4*

- Hosmer, D. et Lemeshow S. Applied Logistic Regression, Wiley.
  $\longrightarrow$ *La régression logistique en applications, en long et en large*

1. Introduction
   - Préambule
   - Analyse bivariée (rappels)
     - Variable quantitative/ Variable quantitative
     - Variable qualitative/ Variable qualitative
     - Variable quantitative/ Variable qualitative
   - Aspects généraux sur la modélisation

Type de variables

On s'intéresse au lien entre 2 variables X et Y.

On distingue deux grandes catégories, chacune déclinées en deux types.

• **Variable quantitative** : son observation est une quantité mesurée.
Exemples : âge, salaire, nombre d'infractions,...

On distingue les variables quantitatives **discrètes** dont les valeurs possibles sont finies ou dénombrables (*Exemples : nombre d'enfants, nombre d'infractions,...*) et les variables quantitatives **continues** qui peuvent prendre toutes les valeurs possibles d'un intervalle (*Exemples : taille, salaire,...*)

• **Variable qualitative** (ou **facteur**) : son observation se traduit par une catégorie ou un code. Les observations possibles sont appelées les **modalités** de la variable qualitative.
Exemples : sexe, CSP, nationalité, mention au BAC,...

Lorsqu'un ordre naturel apparaît dans les modalités, on parle de variable qualitative **ordinale** (*Exemples : mention au BAC,...*). Dans le cas contraire on parle de variable qualitative **nominale** (*Exemples : sexe, CSP,...*).

Exemple du jeu de données "Pottery" : Composition chimique de poteries trouvées sur différents sites archéologiques au Royaume Uni.

	Site	Al	Fe	Mg	Ca	Na
1	Llanedyrn	14.4	7.00	4.30	0.15	0.51
2	Llanedyrn	13.8	7.08	3.43	0.12	0.17
3	Llanedyrn	14.6	7.09	3.88	0.13	0.20
4	Llanedyrn	10.9	6.26	3.47	0.17	0.22
5	Caldicot	11.8	5.44	3.94	0.30	0.04
6	Caldicot	11.6	5.39	3.77	0.29	0.06
7	IsleThorns	18.3	1.28	0.67	0.03	0.03
8	IsleThorns	15.8	2.39	0.63	0.01	0.04
9	IsleThorns	18	1.88	0.68	0.01	0.04
10	IsleThorns	20.8	1.51	0.72	0.07	0.10
11	AshleyRails	17.7	1.12	0.56	0.06	0.06
12	AshleyRails	18.3	1.14	0.67	0.06	0.05
13	AshleyRails	16.7	0.92	0.53	0.01	0.05

Les individus : les poteries numérotées de 1 à 13
Les variables : le site archéologique (facteur à 4 modalités) et différents composés chimiques (quantitatives).

Exemple du jeu de données "NO2trafic" : Concentration en NO2 mesurée à l'intérieur de voitures circulant en région parisienne, selon le type de voie empruntée (5 possibilités) et la fluidité du trafic (de A à D)

	NO2	type	fluidite
1	378.94	P	A
2	806.67	T	D
3	634.58	A	D
4	673.35	T	C
5	589.75	P	A
⋮	⋮	⋮	⋮
283	184.16	P	B
284	121.88	V	D
285	152.39	U	A
286	129.12	U	C

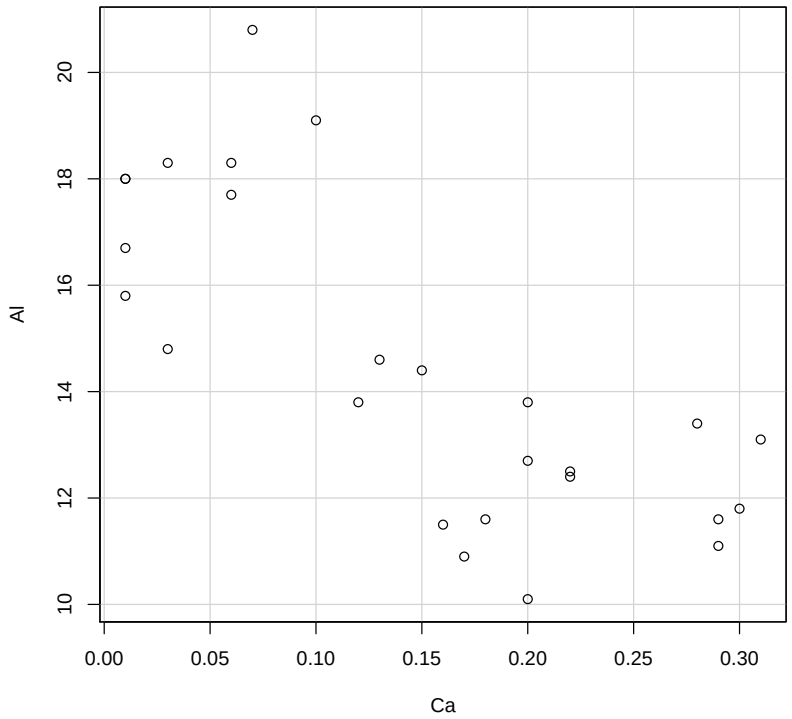
Les individus : les véhicules numérotées de 1 à 286
Les variables : NO2 (quantitative), type (facteur à 5 modalités) et fluidite (facteur ordinal à 4 modalités)

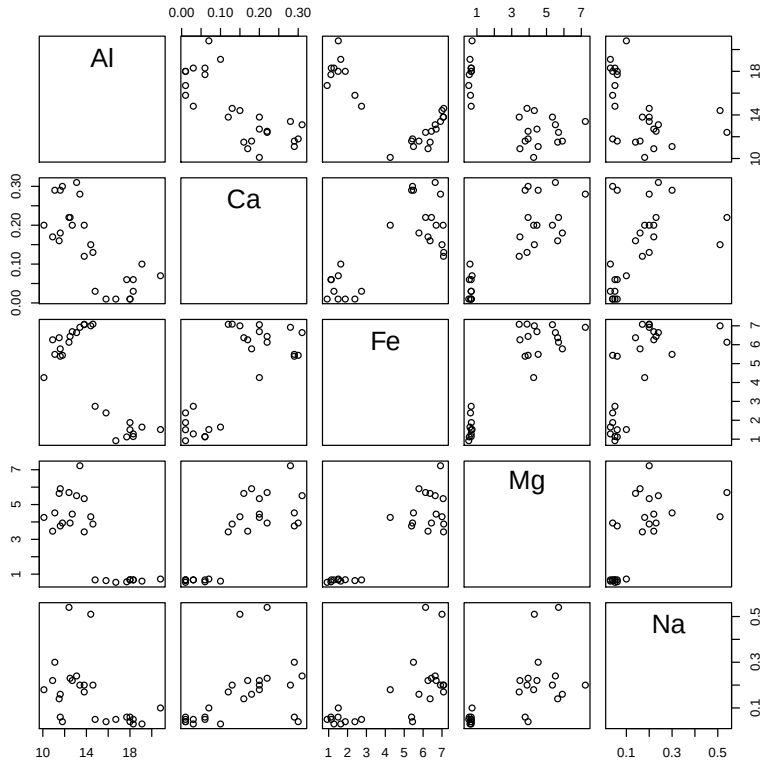
Représentation graphique : nuage de points

Soit $x_1, \ldots, x_n$ les valeurs observées de la première variable quantitative X.
Soit $y_1, \ldots, y_n$ les valeurs observées de la seconde variable quantitative Y.

On visualise le lien entre X et Y grâce au nuage des points (x_i, y_i).

Exemple : nuage de points entre "Al" et "Ca" des données "Pottery" et matrice des nuages de points entre toutes les variables.





Le lien linéaire est quantifié par la **corrélation linéaire** de Pearson :

$$\hat{\rho} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

où *var* et *cov* désignent la variance et la covariance empirique.

Propriétés : On déduit de l'inégalité de Cauchy Schwartz que

- La corrélation $\hat{\rho}$ est toujours comprise entre -1 et 1 :
- si $\hat{\rho} = 1$, il y a un lien linéaire "parfait" positif, i.e. :

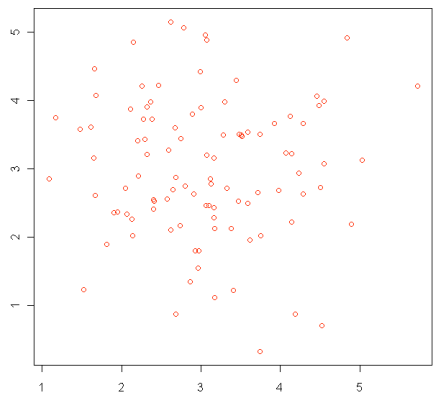
  $\hat{\rho} = 1 \quad$ ssi $\quad$ il existe $a > 0$ et b tel que $y_i = ax_i + b$ pour tout $i = 1, \ldots, n$

- si $\hat{\rho} = -1$, il y a un lien linéaire "parfait" négatif, i.e. :

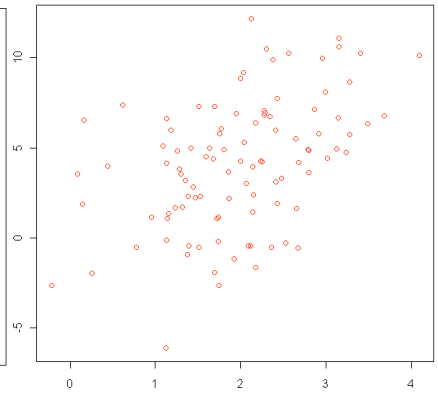
  $\hat{\rho} = -1 \quad$ ssi $\quad$ il existe $a < 0$ et b tel que $y_i = ax_i + b$ pour tout $i = 1, \ldots, n$

- si $\hat{\rho} = 0$, il n'y a aucun lien linéaire (mais il peut exister un lien non-linéaire).

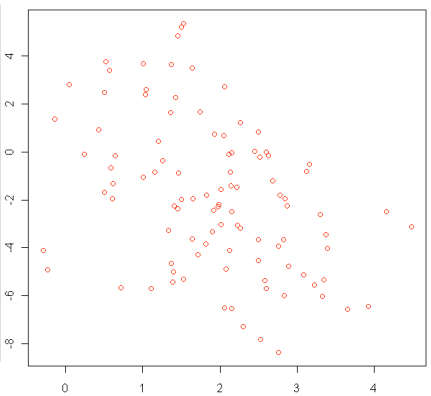
Quelques exemples de nuages de points avec la corrélation correspondante.



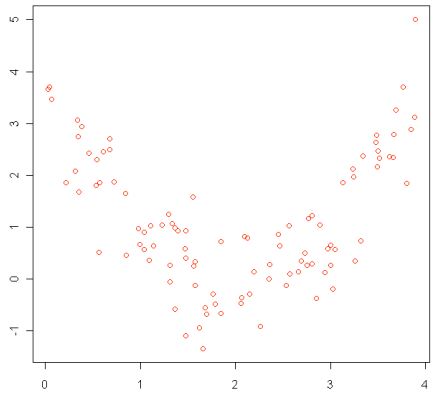
Aucun lien ($\hat{\rho} \approx 0$)



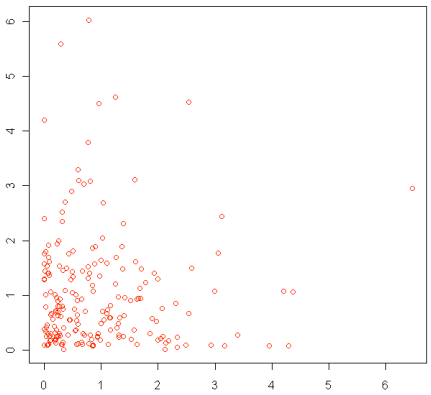
Lien linéaire ($\hat{\rho} \approx 0.4$)



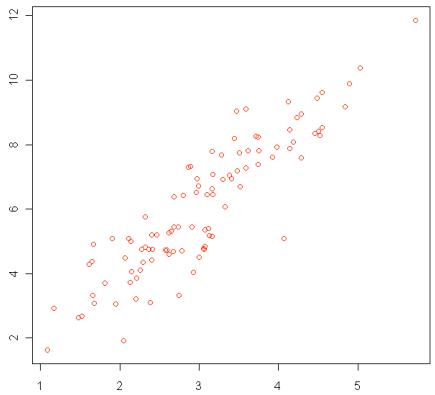
Lien linéaire ($\hat{\rho} \approx -0.4$)



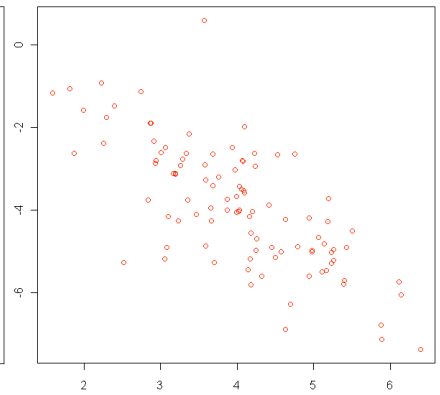
Lien non-linéaire($\hat{\rho} \approx 0$)



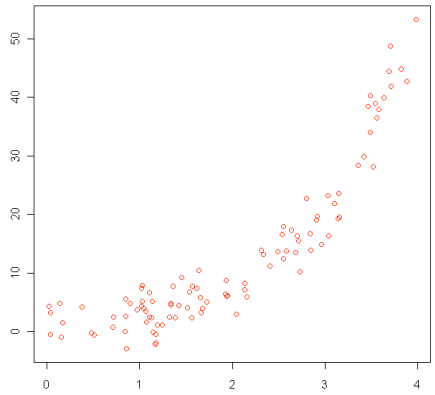
Aucun lien ($\hat{\rho} \approx 0$)



Lien linéaire ($\hat{\rho} \approx 0.9$)



Lien linéaire ($\hat{\rho} \approx -0.8$)



Lien non-linéaire ($\hat{\rho} \approx 0.8$)

Tester si la corrélation est significative

$\hat{\rho}$ est un estimateur de la corrélation théorique ρ entre X et Y défini par

$$\rho = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)]}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

On peut vouloir tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$

Si (X, Y) est Gaussien, on peut montrer que $T \sim St(n-2)$ sous H_0 où

$$T = \sqrt{n-2}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$$

et $St(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté. On en déduit

$$RC_\alpha = \{|T| > t_{n-2}(1-\alpha/2)\}.$$

La p-value vaut quant à elle :

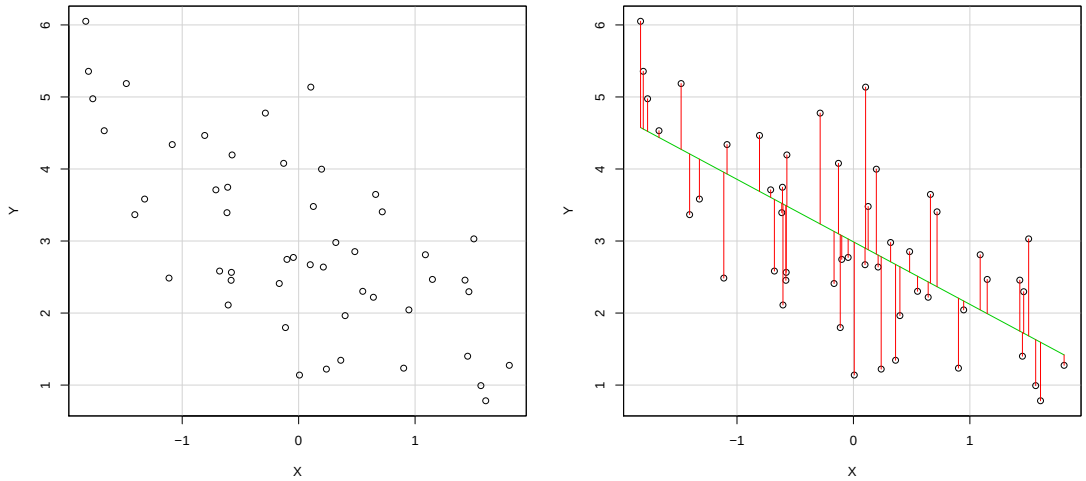
$$p.value = 2\min(F(T), 1 - F(T)),$$

où F est la fonction de répartition de la loi $St(n-2)$.

Sous R : fonction cor.test

Ajuster la droite des moindres carrés

Droite des moindres carrés : Il s'agit de la droite qui passe "le mieux" au milieu des points (x_i, y_i), au sens où la somme des distances en rouge prises au carré est minimale. Il s'agit de la **régression linéaire** de Y sur X.



L'équation de la droite recherchée est donc $y = \hat{a}x + \hat{b}$ où $\hat{a}$ et $\hat{b}$ vérifient :

$$(\hat{a}, \hat{b}) = \underset{(a,b)}{argmin} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

On trouve (à savoir !), si $var(X) \neq 0$:

$$\hat{a} = \frac{cov(X,Y)}{var(X)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

en notant *var* et *cov* la variance et la covariance empirique.

X : premier facteur à I modalités
Y : second facteur à J modalités.
n_{ij} : nombre d'individus ayant la modalité i pour X et j pour Y.
$n_{i.}$: nombre d'individus ayant la modalité i pour X
$n_{.j}$: nombre d'individus ayant la modalité j pour Y

$$n_{i.} = \sum_{j=1}^{J} n_{ij}, \quad n_{.j} = \sum_{i=1}^{I} n_{ij}, \quad n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

Les effectifs n_{ij} sont résumés dans un **tableau de contingence**.

Exemple : Pour les variables "type" et "fluidite" du jeu de données NO2trafic, le tableau de contingence est :

	type				
fluidite	P	U	A	T	V
A	21	21	19	9	9
B	20	17	16	8	7
C	17	17	16	8	7
D	20	20	18	8	8

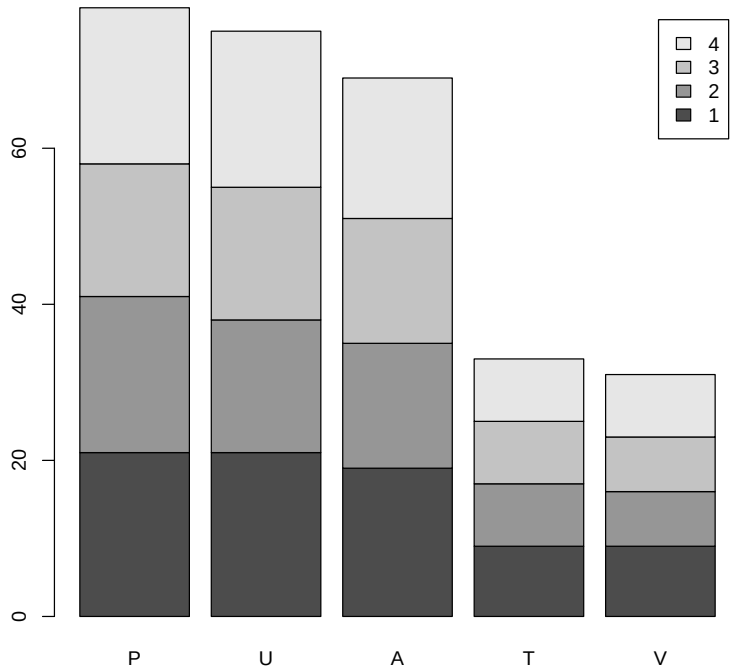
Sous R : table(X,Y)

On résume le tableau de contingence par des diagrammes en batons "croisés", soit par empilement (à gauche), soit côte à côte (à droite).

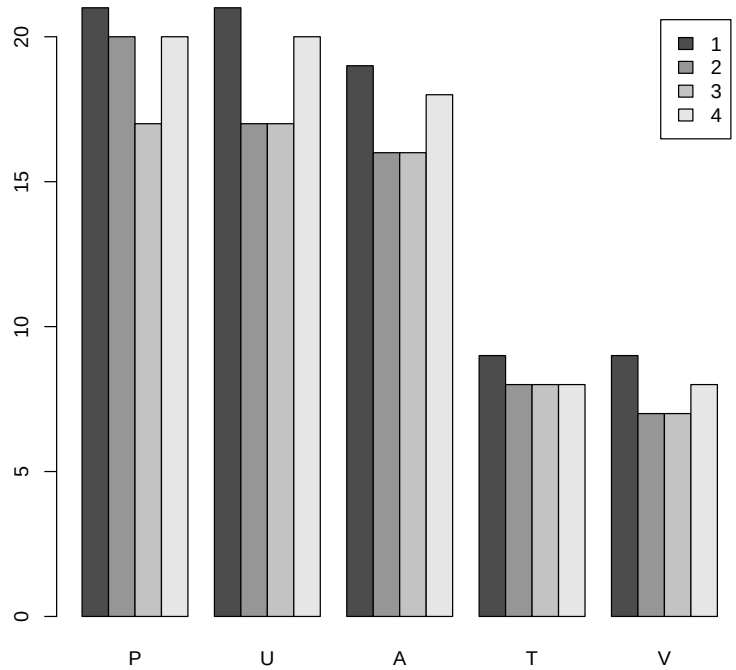
Sous R : si le tableau de contingence se nomme `tab`, barplot(tab) ou barplot(tab,beside=TRUE)

Exemple : pour le graphe croisant les variables "type" et "fluidite",

barplot(tab,legend.text=TRUE)



barplot(tab,beside=TRUE,legend.text=TRUE)



Remarque : si on souhaite représenter les fréquences et non les effectifs, il suffit de diviser `tab` par l'effectif total n, barplot(tab/n).

Pour quantifier le lien entre les deux facteurs, on calcule la distance du χ^2 (khi-deux)

$$d^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cette distance mesure la différence entre les effectifs observés n_{ij} et les effectifs théoriques s'il y avait indépendance : dans ce cas la fréquence observée dans i et j, $\frac{n_{ij}}{n}$, vaudrait le produit des fréquences marginales $\frac{n_{i.}}{n}\frac{n_{.j}}{n}$.

Test du χ^2 : H_0 : X et Y indépendants contre H_1 : le contraire
Sous H_0, $d^2 \sim \chi^2((I-1)(J-1))$ lorsque $n \to \infty$ d'où

$$RC_\alpha = \{d^2 > \chi^2_{(I-1)(J-1)}(1-\alpha)\}$$

est une région critique au niveau asymptotique α, avec $\chi^2_{(I-1)(J-1)}(1-\alpha)$ le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

Sous R : fonction chisq.test

$\to$ Pour aller plus loin dans la compréhension du lien : AFC.

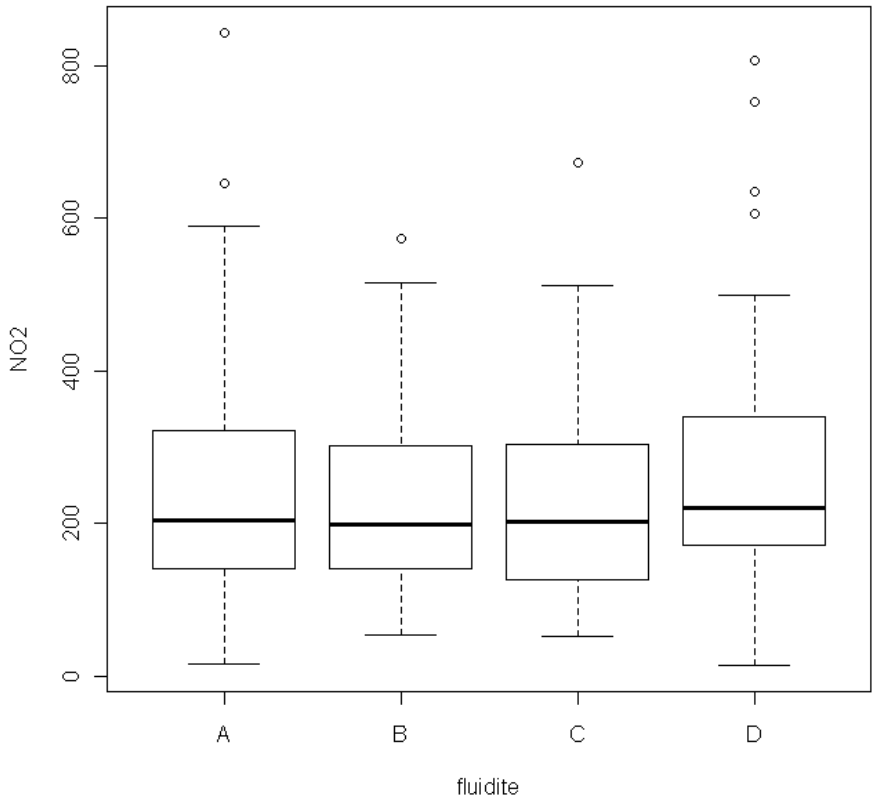
Variable quantitative/ Variable qualitative : graphes

Y : variable quantitative
X : facteur à I modalités

Graphiquement, on effectue des boxplots de Y par modalité de X.

Sous R : boxplot(Y∼ X)

Exemple : dans "NO2trafic", la concentration NO2 en fonction de "fluidite"



Y : variable quantitative
X : facteur à I modalités contenant chacune n_i individus ($\sum_{i=1}^{I} n_i = n$).
y_{ij} : valeur de Y pour l'individu j se trouvant dans la modalité i de X.
On note $\bar{y}_i$ la moyenne de Y dans la modalité i et $\bar{y}$ la moyenne totale, i.e.

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{y}_i$$

Formule de décomposition de la variance : La variance totale est la somme de la variance inter-modalités et de la variance intra-modalités, ce qui s'écrit :

$$\frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{S_T^2} = \frac{1}{n}\underbrace{\sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2}_{S_{inter}^2} + \frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{S_{intra}^2}$$

autrement dit $S_T^2 = S_{inter}^2 + S_{intra}^2$.

Le lien entre X et Y est parfois mesuré par le **rapport de corrélation eta** :

$$\hat{\eta}^2 = \frac{S_{inter}^2}{S_T^2} = \frac{\sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}.$$

On a ainsi $0 \leq \hat{\eta}^2 \leq 1$.

Le coefficient $\hat{\eta}^2$ estime son équivalent théorique η^2 défini par

$$\eta^2 = \frac{\mathbb{V}(\mathbb{E}(Y|X))}{\mathbb{V}(Y)}.$$

Test d'analyse de la variance

En notant $\mu_i = \mathbb{E}(Y|X = i)$ pour $i = 1, \ldots, I$, on souhaite tester

$$H_0 : \mu_1 = \cdots = \mu_I \qquad (\Leftrightarrow \eta^2 = 0)$$

contre H_1 : Y est différent (en espérance) dans au moins deux modalités de X.

Si les y_{ij} sont issus d'une loi Gaussienne de même variance pour tout i, j, alors

$$F = \frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)} = \frac{\hat{\eta}^2/(I-1)}{(1-\hat{\eta}^2)/(n-I)} \sim F(I-1, n-I) \quad \text{sous } H_0.$$

D'où la région critique au niveau α

$$RC_\alpha = \{F > f_{I-1,n-I}(1-\alpha)\}$$

où $f_{I-1,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi $F(I-1, n-I)$.

Sous R : fonction aov(Y∼X) pour obtenir S^2_T, S^2_{inter} et S^2_{intra}, et summary du résultat pour effectuer le test.
Pour $I = 2$, cela correspond au test de Student d'égalité des moyennes (t.test).

Y : grandeur à expliquer (variable d'intérêt)
$X = (X^{(1)}, \ldots, X^{(p)})$: vecteur de p variables explicatives (régresseurs)

But : expliquer/approcher/prédire au mieux Y en fonction de X.

Observations : n individus.
Pour chaque individu $i = 1, \ldots, n$, on observe

Y_i : la valeur de Y correspondant à cet individu
$X_i^{(1)}, \ldots, X_i^{(p)}$: la valeur des $X^{(1)}, \ldots, X^{(p)}$ correspondant à cet individu

Où est l'alea ?

• Généralement, on ne connait aucune valeur *a priori*.
Exemples :
- caractéristiques individuelles d'un client
- caractéristiques environnementales

Dans ce contexte, toutes les variables Y et $X^{(1)}, \ldots, X^{(p)}$ sont aléatoires.
Y_i et $X_i^{(1)}, \ldots, X_i^{(p)}$ en sont les réalisations auprès de chaque individu i.

• Parfois on choisit *a priori* les valeurs de $X = (X^{(1)}, \ldots, X^{(p)})$.
Exemples :
- X : dosages de médicaments (et Y : une réponse physiologique)
- X : dosages de produits phytosanitaires (et Y : productivité d'une culture)

Dans ce contexte, Y est aléatoire, mais pas les $X^{(1)}, \ldots, X^{(p)}$.

Bilan :

- Y est toujours vue comme une variable aléatoire
- $X^{(1)}, \ldots, X^{(p)}$ sont vues comme des variables aléatoires ou des variables déterministes, selon le contexte.

Quel lien cherche-t-on à établir ?

But : approcher au mieux Y en fonction de $X = (X^{(1)}, \ldots, X^{(p)})$.

Mathématiquement : on cherche la "meilleure" fonction f de X qui approche Y.
Meilleure au sens du coût quadratique : on cherche f qui minimise

$$\mathbb{E}[(Y - f(X^{(1)}, \ldots, X^{(p)}))^2].$$

La solution est connue, il s'agit de :

$$f(X^{(1)}, \ldots, X^{(p)}) = \mathbb{E}(Y|X^{(1)}, \ldots, X^{(p)}).$$

Statistiquement : on cherche donc à estimer $f(x) = \mathbb{E}(Y|X = x)$ (pour tous les x pertinents) à partir des n réalisations du couple (Y, X).
Il s'agit de l'objectif de la plupart des modèles de machine learning.

Mais... estimer une fonction de p variables est un peu ambitieux.

$\Longrightarrow$ on fait généralement des hypothèses sur la forme de $\mathbb{E}(Y|X = x)$.

Exemple 1. Si on suppose que la loi de (Y, X) est Gaussienne. Alors on sait que

$$\mathbb{E}(Y|X) = \mathbb{E}(Y) + (X - \mathbb{E}(X))'\beta$$

où $\beta = \Sigma^{-1}(\mathrm{Cov}(Y, X^{(1)}), \ldots, \mathrm{Cov}(Y, X^{(p)}))'$ et où $\Sigma = \mathbb{V}(X)$.

Autrement dit, dans le cas Gaussien,

$$f(X^{(1)}, \ldots, X^{(p)}) = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$$

en notant $\beta = (\beta_1, \ldots, \beta_p)$ et $\beta_0 = \mathbb{E}(Y) - \mathbb{E}(X)'\beta$.

La fonction recherchée f est simplement une fonction affine en $X^{(1)}, \ldots, X^{(p)}$.

- Les paramètres β_0, $\beta_1, \ldots, \beta_p$ sont inconnus en pratique
- Mais le problème d'estimation devient beaucoup plus simple : estimer la fonction f dans ce contexte revient simplement à estimer ces paramètres.

Exemple 2.

- Plutôt que de faire une hypothèse sur la loi de (Y, X), on peut faire une hypothèse sur la forme de $f(X^{(1)}, \dots, X^{(p)}) = \mathbb{E}(Y|X^{(1)}, \dots, X^{(p)})$.
- Si on réduit l'espace des possibles pour f, on facilite son estimation.
- Hypothèse la plus simple : f est affine

$$f(X^{(1)}, \dots, X^{(p)}) = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$$

  pour certains paramètres $\beta_0, \beta_1, \dots, \beta_p$, qu'il conviendra d'estimer.
- C'est la forme exacte de f lorsque (Y, X) est Gaussien (Exemple 1).
- C'est une approximation, plus ou moins bonne, dans les autres cas.
- Il s'agit du cadre de la **régression linéaire**.

Exemple 3. Dans certains cas, f est contraint

- Si Y ne prend que 2 valeurs (disons 0 et 1), alors

$$\mathbb{E}(Y|X^{(1)},\dots,X^{(p)}) = \mathbb{P}(Y=1|X^{(1)},\dots,X^{(p)})$$

- C'est une proba : f doit donc être à valeurs dans $[0,1]$
- Dans ce cas, une forme affine n'est pas adaptée (car à valeurs dans $\mathbb{R}$)
- Par contre, il est possible de supposer une forme affine à une transformation de $\mathbb{E}(Y|X^{(1)},\dots,X^{(p)})$.
- Soit g une fonction (de notre choix) qui va de $[0,1]$ dans $\mathbb{R}$, on peut supposer que $g(f(X^{(1)},\dots,X^{(p)}))$ est affine.
- La modélisation a du sens et requiert peu de paramètres à estimer.
- Cette approche rentre dans le cadre des **modèles linéaires généralisés**.

On a vu que l'objectif général est d'estimer

$$f(X^{(1)}, \ldots, X^{(p)}) = \mathbb{E}(Y|X^{(1)}, \ldots, X^{(p)}).$$

Ce point de vue suppose implicitement :

- que les $X^{(1)}, \ldots, X^{(p)}$ sont des variables aléatoires,
- que toutes les variables sont quantitatives.

Que se passe-t-il sinon ?

Compléments : cas des $X^{(1)}, \ldots, X^{(p)}$ déterministes

Supposons que Y est aléatoire mais pas les $X^{(1)}, \ldots, X^{(p)}$.

- Dans ce cas, le but reste d'expliquer Y au mieux en fonction des $X^{(1)}, \ldots, X^{(p)}$ (déterministes).
- On suppose que la loi de Y dépend de $X^{(1)}, \ldots, X^{(p)}$.
- Sinon l'objectif initial n'a pas de sens.
- En particulier, l'espérance de Y dépend de $X^{(1)}, \ldots, X^{(p)}$, i.e.

$$\mathbb{E}(Y) = f(X^{(1)}, \ldots, X^{(p)}).$$

- Et on cherche à estimer f : l'objectif est donc similaire au cas précédent.
- Dans les deux cas, on cherche à estimer l'espérance de Y **sachant** les variables $X^{(1)}, \ldots, X^{(p)}$.
- Les méthodes d'estimation sont identiques.

On encode chaque variable qualitative à l'aide d'indicatrices.
On parle de "one hot encoding" en machine learning.

• Supposons que Y soit une variable qualitative à 2 modalités ("A" ou "pas A"). Pour modéliser Y, on modélise $\tilde{Y} = \mathbb{1}_{Y="A"}$, c'est à dire

$$\tilde{Y} = \begin{cases} 1 \text{ si } Y = "A" \\ 0 \text{ sinon.} \end{cases}$$

Cette variable binaire peut alors se modéliser comme dans l'exemple 3.

• De même, si Y prend K modalités $A_1, \ldots, A_K$, on introduit les K variables $\tilde{Y}_k = \mathbb{1}_{Y="A_k"}$. Modéliser Y dans ce cas revient donc à modéliser K variables binaires. En réalité seulement $K-1$ car la dernière modalité se déduit des autres ($\sum_{k=1}^{K} \tilde{Y}_k = 1$).

• On applique si besoin la même transformation aux variables $X^{(1)}, \ldots, X^{(p)}$. Cette transformation amène toutefois quelques spécificités dans l'écriture du modèle et son interprétation : cf le chapitre sur l'ANOVA et l'ANCOVA.

On suppose dans ce chapitre que :
Y : variable **quantitative**
$X^{(1)}, \ldots, X^{(p)}$: p variables **quantitatives**

Pour rappel, on cherche à trouver f telle que $Y \approx f(X^{(1)}, \ldots, X^{(p)})$.
La fonction idéale est

$$f(X^{(1)}, \ldots, X^{(p)}) = \mathbb{E}(Y | X^{(1)}, \ldots, X^{(p)}).$$

On observe, pour chaque individu $i = 1, \ldots, n$, Y_i et $X_i^{(1)}, \ldots, X_i^{(p)}$.

Mais estimer f en tout généralité à partir de ces observations est trop ambitieux.

En **régression linéaire** on suppose que

$$\mathbb{E}(Y|X^{(1)},\dots,X^{(p)}) = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$$

pour certains paramètres $\beta_1,\dots,\beta_p$ inconnus.

Cette hypothèse revient à dire que pour chaque individu $i = 1,\dots,n$

$$Y_i = \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \varepsilon_i$$

où ε_i vérifie $\mathbb{E}(\varepsilon_i|X^{(1)},\dots,X^{(p)}) = 0$.

A partir des observations, on souhaite :

- estimer les paramètres $\beta_1,\dots,\beta_p$,
- valider la relation précédente.

Pour simplifier les écritures, on va supposer dans tout le chapitre que les variables $X^{(1)}, \dots, X^{(p)}$ ne sont pas aléatoires.
(Sinon, il faut ajouter des "sachant $X^{(1)}, \dots, X^{(p)}$" dans toutes les formules.)

Le modèle s'écrit donc, pour chaque individu $i = 1, \dots, n$,

$$Y_i = \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \varepsilon_i$$

où ε_i vérifie $\mathbb{E}(\varepsilon_i) = 0$.

Remarque :

La première variable vaut généralement $X_i^{(1)} = 1$ pour tout i.
Le modèle est alors affine en $X^{(2)}, \dots, X^{(p)}$.
On parle alors d'un modèle *avec constante* (elle vaut β_1).
Certains ouvrages notent cette constante β_0.
Le fait d'introduire $X^{(1)}$ permet une présentation unifiée.

On regroupe les observations dans les vecteurs

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X^{(1)} = \begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_n^{(1)} \end{pmatrix}, \quad \cdots \quad , X^{(p)} = \begin{pmatrix} X_1^{(p)} \\ \vdots \\ X_n^{(p)} \end{pmatrix}.$$

On introduit la matrice X de taille (n, p) regroupant toutes les variables explicatives, appelée également "matrice de design" :

$$X = (X^{(1)}|\dots|X^{(p)}) = \begin{pmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \vdots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{pmatrix}.$$

On note enfin :

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Le modèle sur les n observations s'écrit alors :

$$\boxed{Y = X\beta + \varepsilon}$$

Petite mise au point :

La notation X a été utilisée pour plusieurs objets :

- une variable aléatoire ;
- un vecteur de variables aléatoires ;
- une matrice.

Il n'y a normalement pas ambiguïté dans le contexte.

Dans la suite du chapitre, X sera toujours la matrice du slide précédent.

De même, Y sera le vecteur du slide précédent.

Avec ces notations, les hypothèses du modèle linéaire sont :

$$\underset{(n,1)}{Y} = \underset{(n,p)}{X}\underset{(p,1)}{\beta} + \underset{(n,1)}{\varepsilon}$$

(en indiquant les dimensions) avec

1. $\mathbb{E}(\varepsilon) = 0$.
2. $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, où I_n est la matrice identité.
3. $rg(X) = p$: les vecteurs $X^{(1)}, \ldots, X^{(p)}$ sont linéairement indépendants.

(Rappel : on a supposé X non aléatoire pour simplifier les écritures)

Commentaires :

1. signifie qu'on a bien modélisé l'espérance de Y
2. est une double hypothèse :
   - homoscédasticité (la variance est la même pour tous les individus)
   - non-corrélation des individus entre eux
3. Si une variable $X^{(j)}$ était combinaison linéaire des autres, alors le modèle ne serait pas identifiable (une infinité de paramètres β conduiraient au même modèle – illustration au tableau –).
   Cela implique également $p \leq n$. Le cas $p > n$ nécessite des méthodes spécifiques de "régression en grande dimension".

Exemple de la régression simple

On parle de régression simple lorsqu'il n'y a qu'un seul régresseur.
On observe donc le couple (x_i, y_i) pour $i = 1, \dots, n$, et le modèle s'écrit

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Avec les notations précédentes, $p = 2$, $X_i^{(1)} = 1$ et $X_i^{(2)} = x_i$.
On a donc $Y = X\beta + \varepsilon$ où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X^{(1)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad X^{(2)} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Autre exemple

Exemple :
Supposons qu'on observe (x_i, y_i) pour $i = 1, \dots, n$.
Le modèle

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$$

est un modèle de régression linéaire.

Ici, $p = 3$, $X_i^{(1)} = 1$, $X_i^{(2)} = x_i$ et $X_i^{(3)} = x_i^2$.

Il ne s'agit pas d'une relation linéaire entre y_i et x_i.
Par contre, les paramètres β_1, β_2 et β_3 interviennent de façon linéaire.

La "linéarité" dans le modèle de régression $Y = X\beta + \varepsilon$ concerne β.
En particulier, il permet de la non-linéarité par rapport aux régresseurs.

2 Régression linéaire

- Modélisation
- Inférence
  - Estimation de β
  - Estimation de σ^2
  - Cas Gaussien
  - Tests et intervalles de confiances pour β_j
  - Prévision
- Validation
- Critères de sélection de modèles

Les MCO consistent à trouver la valeur $\hat{\beta}$ du vecteur β qui minimise

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \beta_1 X_i^{(1)} - \cdots - \beta_p X_i^{(p)} \right)^2 .$$

Théorème

Si $rg(X) = p$,

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Preuve : Il s'agit de projeter de façon orthogonale le vecteur Y sur l'espace vectoriel engendré par les vecteurs $X^{(1)}, \ldots, X^{(p)}$. Si $rg(X) = p$, cette projection est unique. $\hat{\beta}$ correspond alors aux coordonnées du projeté. Cf le tableau pour les détails.

Il n'est pas inutile d'insister : l'estimation par MCO est une **projection**.

Quelques notations utilisées dans la preuve précédente et dans la suite :

- On note $[X]$ l'espace vectoriel engendré par les vecteurs $X^{(1)}, \dots, X^{(p)}$, i.e.,

$$[X] = \{X\alpha,\ \alpha \in \mathbb{R}^p\} = \{v \in \mathbb{R}^n,\ \exists \alpha \in \mathbb{R}^p,\ v = X\alpha\}.$$

- On note $P_{[X]}$ la matrice de projection orthogonale sur $[X]$

$$P_{[X]} = X(X'X)^{-1}X'.$$

- On note $\hat{Y}$ le projeté sur $[X]$.

$$\hat{Y} = P_{[X]}Y = X\hat{\beta} = X(X'X)^{-1}X'Y.$$

- On note $[X]^{\perp}$ l'espace vectoriel orthogonal à $[X]$ dans $\mathbb{R}^n$,

$$[X]^{\perp} = \{v \in \mathbb{R}^n,\ X'v = 0\}$$

- La matrice de projection orthogonale sur $[X]^{\perp}$ est

$$P_{[X]^{\perp}} = I_n - P_{[X]} = I_n - X(X'X)^{-1}X'.$$

Il n'est pas inutile d'insister : l'estimation par MCO est une **projection**.

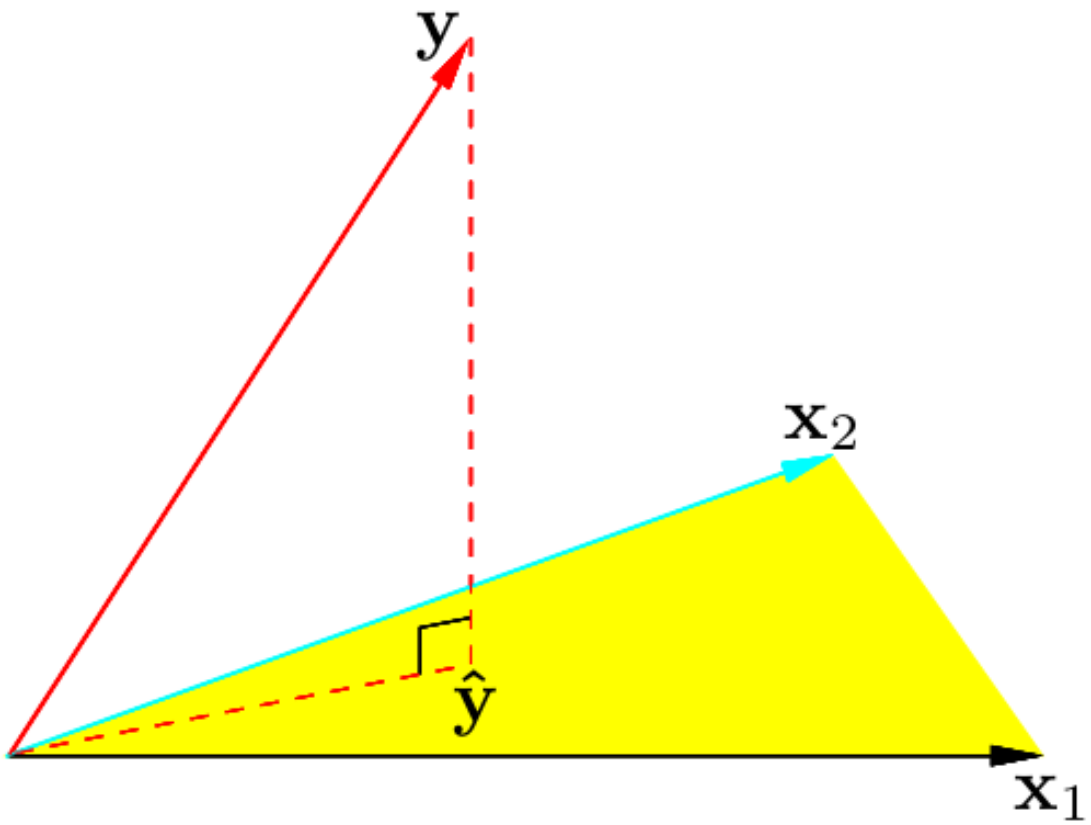


Figure extraite de l'ouvrage ESL. Le plan en jaune représente $[X]$ lorsque $p = 2$.
Le vecteur Y est projeté sur $[X]$ pour donner $\hat{Y} = X\hat{\beta}$.

Quelques propriétés concernant la qualité d'estimation :

Proposition

Si $rg(X) = p$, $\mathbb{E}(\varepsilon) = 0$ *et* $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, *alors*

$$\mathbb{E}(\hat{\beta}) = \beta \quad (\hat{\beta} \text{ est un estimateur sans biais})$$
$$\mathbb{V}(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

Si de plus $(X'X)^{-1}$ *tend vers 0 lorsque* $n \to +\infty$, *alors* $\hat{\beta}$ *converge en moyenne quadratique vers* β.

Preuve : Cf le tableau

Théorème (de Gauss-Markov)

Si $rg(X) = p$, $\mathbb{E}(\varepsilon) = 0$ *et* $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, *alors* $\hat{\beta}$ *est le meilleur estimateur linéaire sans biais de* β, *au sens du coût quadratique.*

Preuve : Cf le tableau

On introduit les **résidus** : pour $i = 1, \dots, n$

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

On note le vecteur des résidus :

$$\hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}.$$

Proposition

- $\hat{\varepsilon} = Y - \hat{Y} = Y - P_{[X]}Y = P_{[X]^\perp}Y = P_{[X]^\perp}\varepsilon$.
- *Si* $rg(X) = p$, $\mathbb{E}(\varepsilon) = 0$ *et* $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, *alors*

$$\mathbb{E}(\hat{\varepsilon}) = 0 \quad \text{et} \quad \mathbb{V}(\hat{\varepsilon}) = \sigma^2 P_{[X]^\perp} = \sigma^2 (I_n - X(X'X)^{-1}X').$$

- *Si le modèle contient une constante, typiquement* $X_i^{(1)} = 1$ *pour tout* i, *alors* $\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = 0$, *ou de manière équivalente* $\bar{\hat{Y}} = \bar{Y}$.

Preuve : Cf le tableau

Proposition

Si $rg(X) = p$, $\mathbb{E}(\varepsilon) = 0$ *et* $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, *alors*

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-p}\|\hat{\varepsilon}\|^2$$

est un estimateur sans biais de σ^2.
Si de plus les ε_i *sont i.i.d, alors* $\hat{\sigma}^2$ *est un estimateur consistant de* σ^2.

Preuve : Cf le tableau

On utilise souvent la notation :

$$\hat{\sigma}^2 = \frac{SCR}{n-p},$$

où $SCR = \|\hat{\varepsilon}\|^2$ est la somme des carrés des résidus.

- Jusqu'alors, on a supposé $Y = X\beta + \varepsilon$ avec $\mathbb{E}(\varepsilon) = 0$ et $\mathbb{V}(\varepsilon) = \sigma^2 I_n$.
- Aucune hypothèse sur la loi de ε (et donc de Y) n'a été formulée.
- Dans cette partie, on suppose que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
- Cela implique que $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.
- Il s'agit d'une hypothèse souvent faite par défaut dans les modèles de régression linéaire.
- Le fait de connaître entièrement la loi de Y ouvre la voie à l'estimation par maximum de vraisemblance.
- Le modèle $Y = X\beta + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et $rg(X) = p$ sera appelé **modèle Gaussien** par la suite.

Proposition

Dans le modèle Gaussien, en notant $\hat{\beta}_{MV}$ et $\hat{\sigma}^2_{MV}$ les estimateurs du maximum de vraisemblance de β et σ^2 (et $\hat{\beta}$ et $\hat{\sigma}^2$ les estimateurs par MCO) on a :

- $\hat{\beta}_{MV} = \hat{\beta}$ *et* $\hat{\sigma}^2_{MV} = \frac{SCR}{n} = \frac{n-p}{n}\hat{\sigma}^2$.
- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$.
- $\frac{n-p}{\sigma^2}\hat{\sigma}^2 = \frac{n}{\sigma^2}\hat{\sigma}^2_{MV} \sim \chi^2(n-p)$.
- $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants*

Preuve : Cf le tableau

Théorème

Dans le modèle Gaussien, $\hat{\beta}$ est un estimateur efficace de β, c'est à dire qu'il s'agit du meilleur estimateur sans biais possible de β.

Preuve : Cf le tableau

Cas Gaussien : remarques

- Ce dernier théorème renforce le théorème de Gauss-Markov : sans l'hypothèse de normalité, $\hat{\beta}$ est optimal parmi les estimateurs *linéaires* sans bias ; avec, $\hat{\beta}$ est optimal parmi *tous* les estimateurs sans biais.

- Dans le cas Gaussien, on connait la loi exacte de $\hat{\beta}$ et $\hat{\sigma}^2$: cela permet de construire des tests et des intervalles de confiance (cf partie suivante)

- Cependant, si le modèle n'est pas Gaussien (et sous quelques conditions faibles), les lois de $\hat{\beta}$ et $\hat{\sigma}^2$ convergent vers les mêmes lois que dans le cas Gaussien, lorsque $n \to \infty$. Il s'agit d'une application (d'une généralisation) du Théorème Limite Central.

- Ainsi, si n est grand (quelques dizaines suffisent), **la plupart des résultats montrés dans le cas Gaussien restent valables dans le cas général**.

On rappelle que $\hat{\sigma}^2 = \frac{1}{n-p}\|\hat{\varepsilon}\|^2$.

Corollaire

Dans le modèle Gaussien, pour tout $j = 1, \ldots, p$,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}} \sim St(n-p)$$

où $(X'X)^{-1}_{jj}$ désigne le j-ème élément de la diagonale de la matrice $(X'X)^{-1}$.

Preuve : Cf le tableau

Remarques

- Le dénominateur n'est autre qu'une estimation de l'écart-type de $\hat{\beta}_j$.
- En effet $\mathbb{V}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ implique que $\mathbb{V}(\hat{\beta}_j) = \sigma^2(X'X)^{-1}_{jj}$.
- On note parfois $\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$.

Ainsi, dans le modèle Gaussien, pour tester

$$H_0 : \beta_j = 0 \quad \text{contre} \quad H_1 : \beta_j \neq 0,$$

pour $j \in \{1, \dots, p\}$, une région critique au niveau α est

$$RC_\alpha = \left\{ \frac{|\hat{\beta}_j|}{\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}} > t_{n-p}(1-\alpha/2) \right\},$$

où $t_{n-p}(1-\alpha/2)$ désigne le quantile d'ordre $1-\alpha/2$ d'une loi $St(n-p)$.

On note parfois la statistique de test $T_j = \hat{\beta}_j / \hat{\sigma}_{\hat{\beta}_j}$. La p-valeur associée à ce test se calcule ainsi :

$$p.value = 2\min(F(T_j), 1 - F(T_j)),$$

où F est la fonction de répartition de la loi $St(n-p)$.

- Ce test est appelé **test de significativité de Student**.
- Il permet en effet de tester si la variable $X^{(j)}$ est *significative*.
- Si $\beta_j = 0$, la variable n'a aucun effet : on peut l'enlever du modèle.

On peut en déduire de même un intervalle de confiance (IC) pour β_j.
Dans le modèle Gaussien, un IC au niveau de confiance $1-\alpha$ pour β_j est

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}; \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}\right].$$

Ou en plus abrégé :

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j \pm t_{n-p}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}\right],$$

en notant $\hat{\sigma}_{\hat{\beta}_j}$ l'estimateur de l'écart-type de $\hat{\beta}_j$.

On vérifie en effet d'après le corollaire précédent que

$$\mathbb{P}(\beta_j \in IC_{1-\alpha}(\beta_j)) = 1-\alpha.$$

Gaussien vs pas Gaussien

- Les tests et ICs présentés ci-dessus sont ceux calculés par les logiciels.
- Stricto sensu, ils supposent que le modèle est Gaussien.
- C'est parceque dans ce cas, on connait parfaitement la loi de tout le monde, quel que soit n.
- Néanmoins, lorsque n est grand (quelques dizaines), ces tests et ICs restent valables dans le cas non Gaussien.
- L'hypothèse Gaussienne doit donc être vue davantage comme une hypothèse de confort théorique, que comme une contrainte pratique.

- On suppose qu'on a estimé β et σ^2 par $\hat{\beta}$ et $\hat{\sigma}^2$ à partir des observations des variables Y et $X^{(1)}, \ldots, X^{(p)}$ auprès de n individus.
- On souhaite prédire Y pour un nouvel individu o, c'est à dire prédire Y_o, connaissant les valeurs $X_o^{(1)}, \ldots, X_o^{(p)}$ des régresseurs pour cet individu.
- On suppose que ce nouvel individu suit exactement le même modèle que les autres individus, associé à une erreur ε_o qui lui est propre.

En notant x_o le vecteur de taille p des régresseurs pour l'individu o :

$$x_o = \begin{pmatrix} X_o^{(1)} \\ \vdots \\ X_o^{(p)} \end{pmatrix},$$

on a donc

$$Y_o = x_o'\beta + \varepsilon_o = \beta_1 X_o^{(1)} + \cdots + \beta_p X_o^{(p)} + \varepsilon_o,$$

où $\mathbb{E}(\varepsilon_o) = 0$, $\mathbb{V}(\varepsilon_o) = \sigma^2$ et $Cov(\varepsilon_o, \varepsilon_i) = 0$ pour tout $i = 1, \ldots, n$.

Etant donné le modèle $Y_o = x_o'\beta + \varepsilon_o$, la prévision naturelle de Y_o est

$$\hat{Y}_o = x_o'\hat{\beta}.$$

Deux erreurs se cumulent dans cette prévision :

1. celle due à l'estimation de β par $\hat{\beta}$.
2. et celle due à "l'oubli" de ε_o,

L'erreur de prévision vaut $Y_o - \hat{Y}_o$. On a :

- $\mathbb{E}(Y_o - \hat{Y}_o) = 0$: l'erreur de prévision est nulle en moyenne.
- $\mathbb{V}(Y_o - \hat{Y}_o) = \sigma^2(x_o'(X'X)^{-1}x_o + 1)$

Sa variance intègre les deux types d'erreur évoquées ci-dessus :

1. celle liée à $\hat{\beta}$: elle est négligeable lorsque n est grand (si $(X'X)^{-1} \to 0$) ;
2. celle liée à ε_o : elle vaut toujours σ^2, cette erreur est incompressible.

Si l'on suppose que le modèle est Gaussien, alors

$$Y_o - \hat{Y}_o \sim \mathcal{N}(0, \sigma^2(x_o'(X'X)^{-1}x_o + 1)).$$

On en déduit que

$$\frac{Y_o - \hat{Y}_o}{\hat{\sigma}\sqrt{x_o'(X'X)^{-1}x_o + 1}} \sim St(n-p)$$

et on peut donc fournir un intervalle de prévision pour Y_o :

$$IP_{1-\alpha}(Y_o) = \left[\hat{Y}_o \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{x_o'(X'X)^{-1}x_o + 1}\right],$$

dans le sens où $\mathbb{P}(Y_o \in IP_{1-\alpha}(Y_o)) = 1 - \alpha$.

Remarques :

- Pour une fois, l'hypothèse Gaussienne est vraiment importante.
- C'est parce qu'on utilise la Gaussianité de ε_o, peu importe la valeur de n.
- C'est le seul endroit du cours où l'hypothèse Gaussienne est importante.

Intervalles de prévision versus intervalles d'estimation

Pour un x_o donné, on peut se poser 2 types de questions :

1. Prédire $Y_o = x_o'\beta + \varepsilon_o$, comme on vient de le faire, par $\hat{Y}_o = x_o'\hat{\beta}$.
   Pour rappel, la variance de l'erreur de prévision inclut :
   - l'erreur due à l'estimation de β,
   - l'oubli de ε_o.
   D'où l'intervalle de **prévision pour $\mathbf{Y_o}$** :

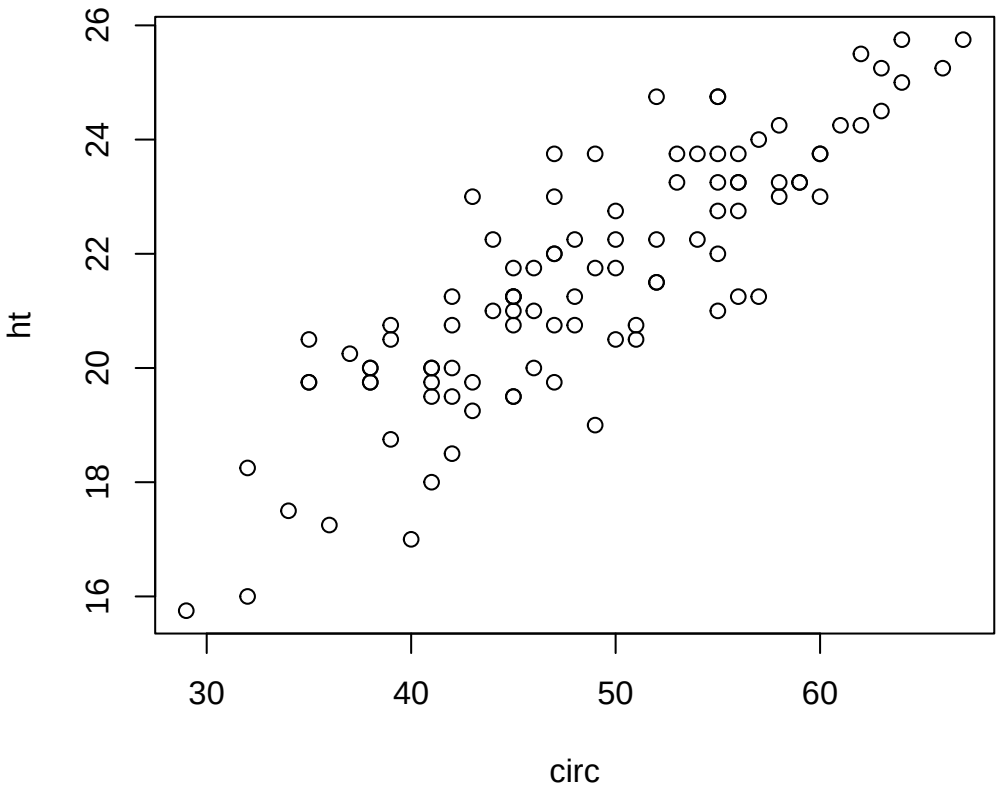
$$IP_{1-\alpha}(Y_o) = \left[\hat{Y}_o \pm t_{n-p}(1-\alpha/2)\sqrt{\hat{\sigma}^2 x_o'(X'X)^{-1}x_o + \hat{\sigma}^2}\right].$$

2. Estimer simplement $\mathbb{E}(Y_o) = x_o'\beta$, c'est à dire la valeur sur la "droite" (l'hyperplan) associée à x_o, sans s'intéresser à Y_o.
   On l'estime évidemment par $x_o'\hat{\beta}$, c'est à dire comme $\hat{Y}_o$.
   Par contre l'erreur d'estimation est uniquement due à l'estimation de β.
   D'où l'intervalle de confiance pour **l'estimation de $\mathbf{x_o'\beta}$** :

$$IC_{1-\alpha}(x_o'\beta) = \left[\hat{Y}_o \pm t_{n-p}(1-\alpha/2)\sqrt{\hat{\sigma}^2 x_o'(X'X)^{-1}x_o}\right].$$

Exemple : hauteur en fonction de la circonférence d'arbres

Pour 100 arbres, on relève leur circonférence (`circ`) et leur hauteur (`ht`)



- On cherche à modéliser la hauteur en fonction de la circonférence.
- Un lien linéaire semble adapté.

Exemple : hauteur en fonction de la circonférence d'arbres

On va donc estimer le modèle $ht_i = \beta_1 + \beta_2 circ_i + \varepsilon_i$, pour $i = 1, \ldots, 100$.

Sous R : on utilise la fonction `lm`, puis `summary` :

```
reg= lm(ht~circ)
summary(reg)
```

On obtient (entre autres) ceci :

```
      Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0646     0.6517   16.98   <2e-16 ***
circ          0.2165     0.0132   16.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.132 on 98 degrees of freedom
```

Exemple : hauteur en fonction de la circonférence d'arbres

Lecture de cette sortie :

- La colonne `Estimate` fournit les paramètres estimés par MCO :

$$\hat{\beta}_1 = 11.0646, \quad \hat{\beta}_2 = 0.2165.$$

- La colonne `Std. Error` fournit l'estimation de l'écart-type des estimateurs :

$$\hat{\sigma}_{\hat{\beta}_1} = 0.6517, \quad \hat{\sigma}_{\hat{\beta}_2} = 0.0132.$$

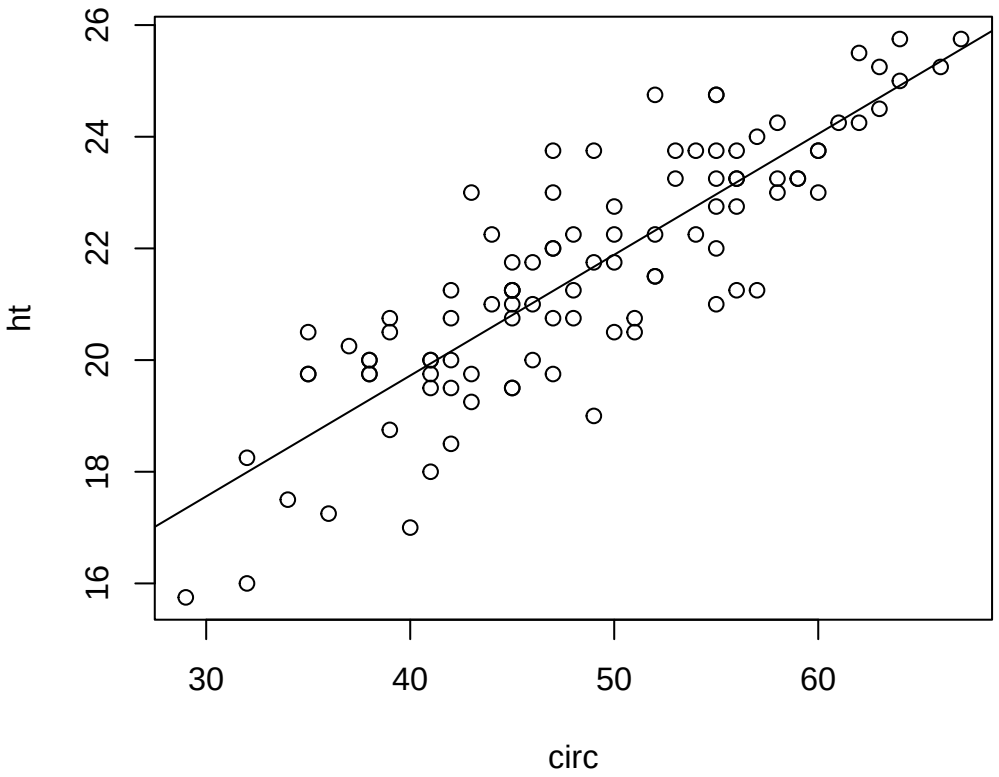
- La colonne `t value` fournit la valeur de la statistique de Student permettant de tester la significativité de chaque variable :

$$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = 16.98, \quad T_2 = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = 16.40.$$

- La colonne `Pr(>|t|)` fournit la p-value du test de significativité : elle est ici négligeable ($< 2.10^{-16}$) pour les deux coefficients.
- Les étoiles résument la significativité de chaque variable, selon la valeur de la p-value : 0 étoile si $p.value > 0.1$ (pas significatif), jusqu'à 3 étoiles si $p.value < 0.001$ (très significatif).
- La dernière ligne fournit $\hat{\sigma} = 1.132$ et le degré de liberté $n - p = 98$.

Exemple : hauteur en fonction de la circonférence d'arbres

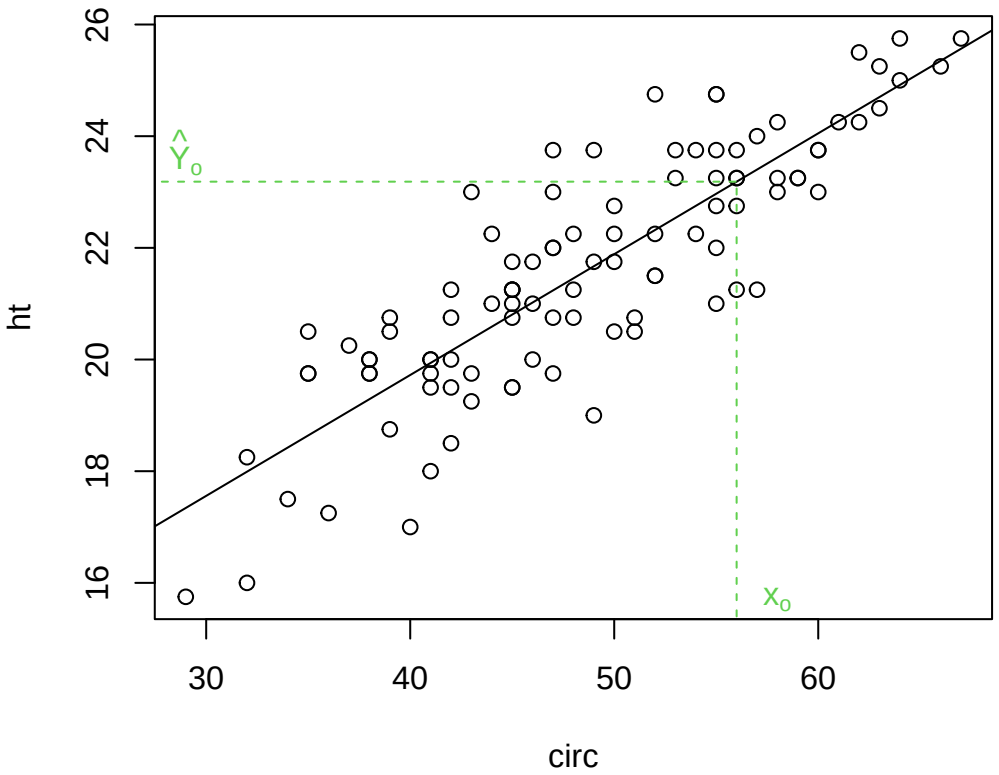
On obtient la droite de régression estimée suivante : $ht = 11.0646 + 0.2165 circ$



Sous R : en notant `reg=lm(ht~circ)` : `abline(reg)`

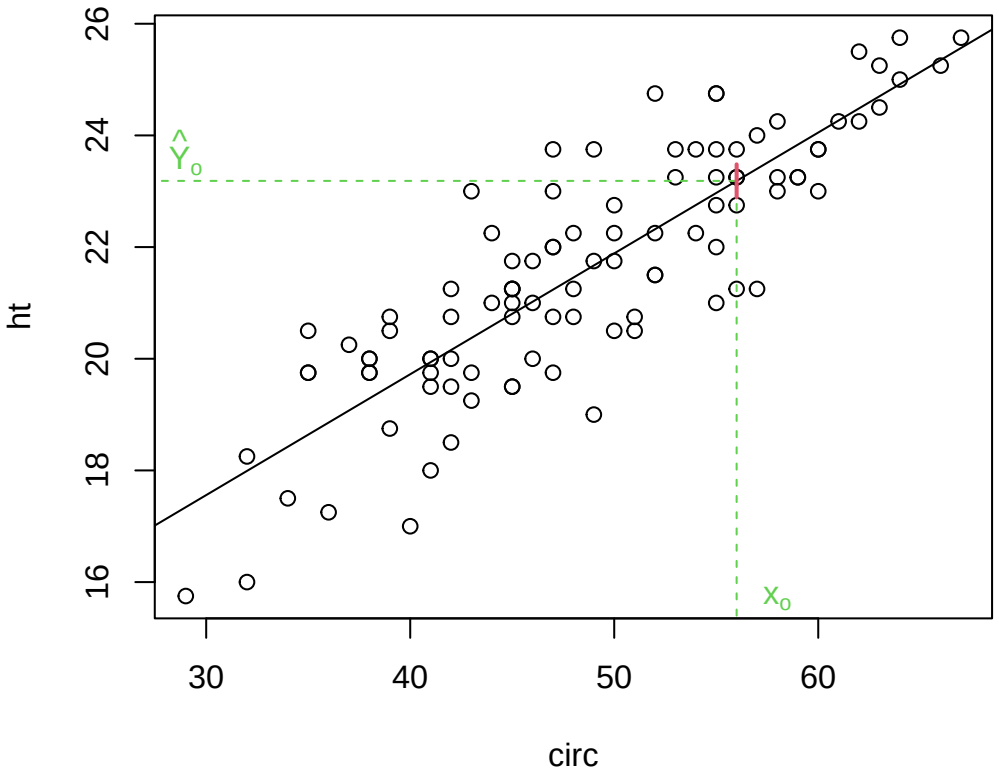
Exemple : hauteur en fonction de la circonférence d'arbres

Pour un x_o donné, on en déduit $\hat{Y}_o$ sur la droite.



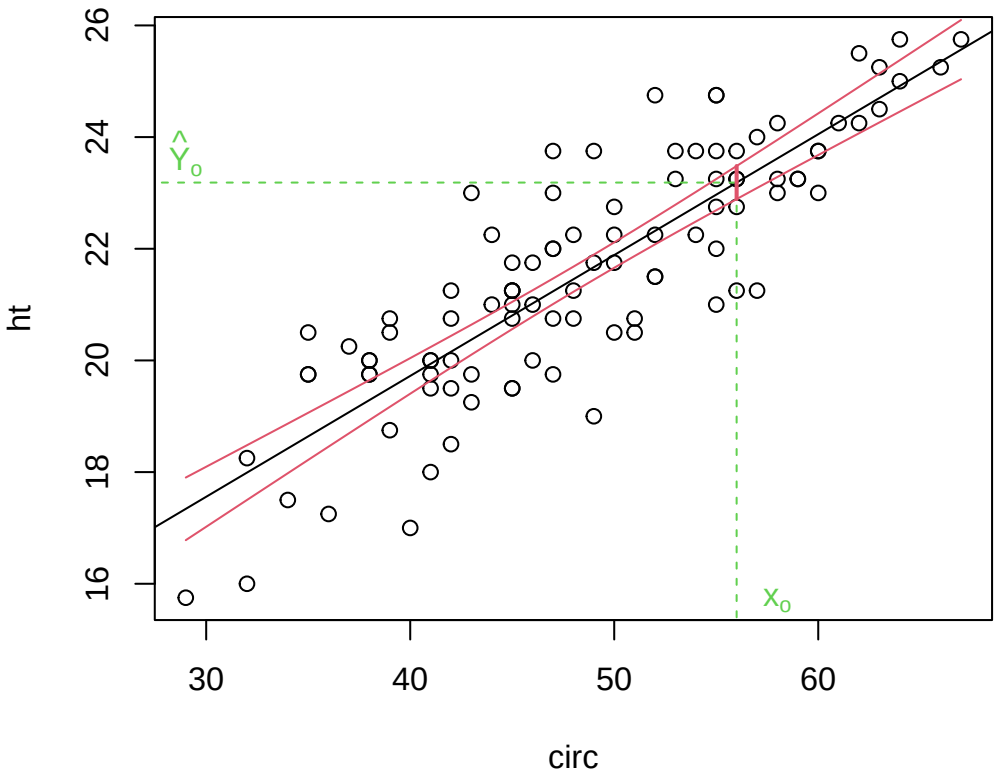
Exemple : hauteur en fonction de la circonférence d'arbres

On peut s'intéresser à **l'intervalle de confiance d'estimation** de la droite en cet abscisse x_o (petit segment rouge).
Il informe sur **où la droite a le plus de chance de se trouver** en cet abscisse.



Exemple : hauteur en fonction de la circonférence d'arbres

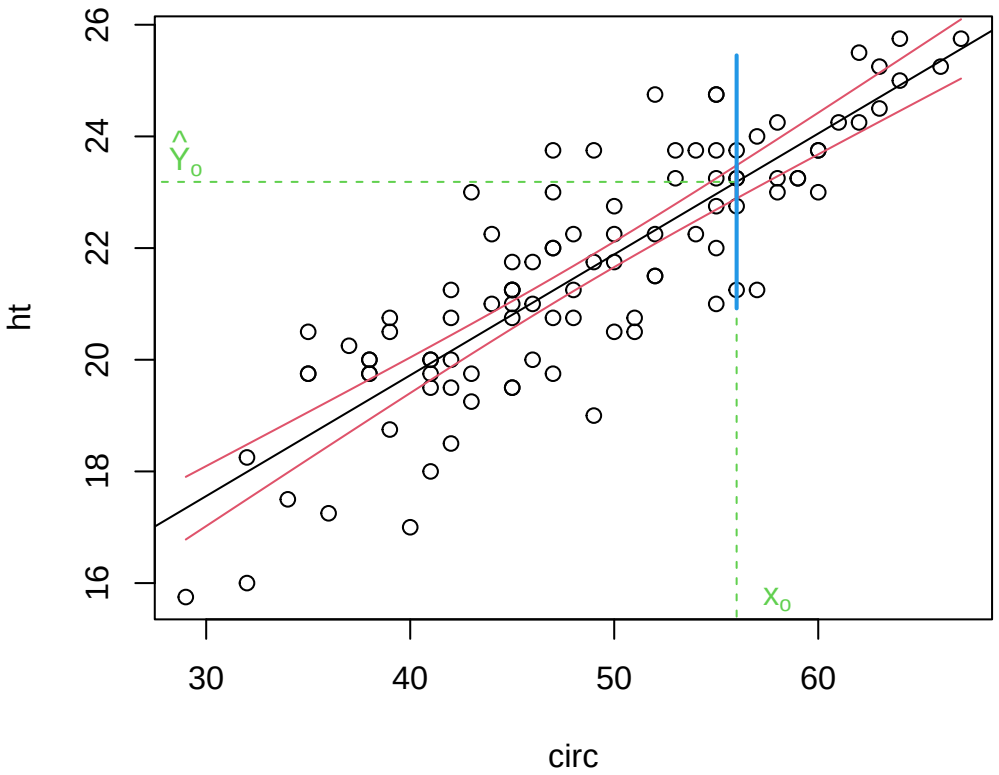
Idem en chaque abscisse x : les courbes rouges représentent l'IC d'estimation de la droite en tout x.



Sous R : `predict(reg,interval="confidence")`

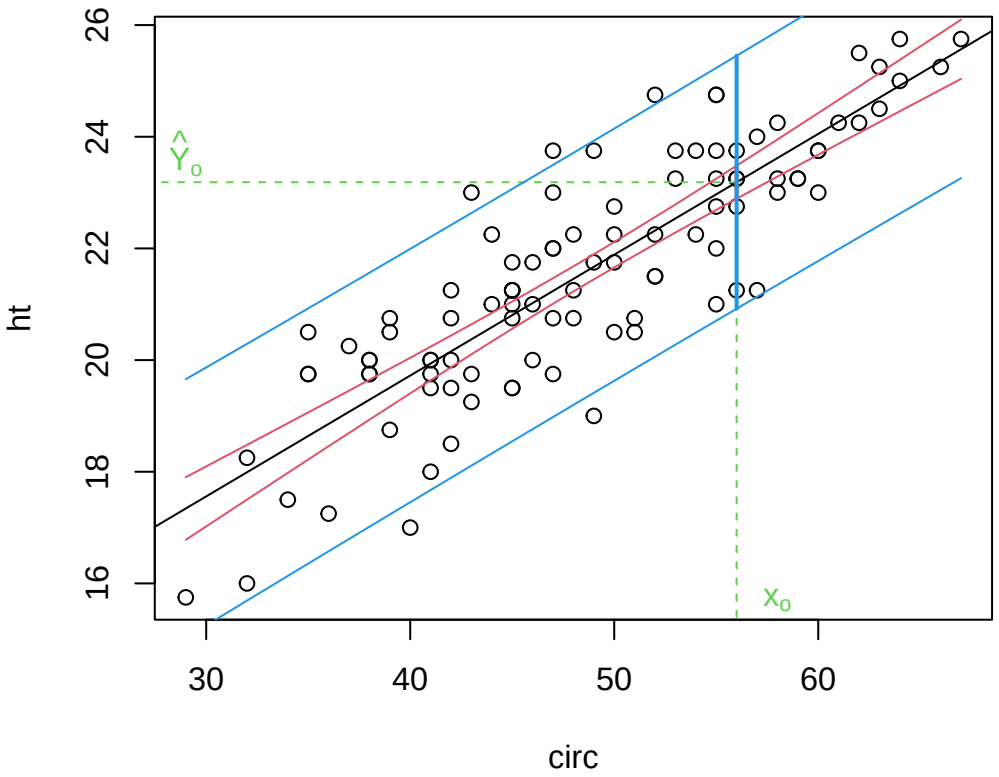
Exemple : hauteur en fonction de la circonférence d'arbres

On peut s'intéresser au contraire à **l'intervalle de prévision** de Y_o (segment bleu).
Il informe sur **où la valeur Y_o a le plus de chance de se trouver**.



Exemple : hauteur en fonction de la circonférence d'arbres

Idem en chaque abscisse x : les courbes bleues représentent l'IP de Y en tout x.



Sous R : `predict(reg,interval="prediction")`

On définit le R^2 à l'aide du théorème de Pythagore. On distingue deux cas :

- Si le modèle contient une constante ($\mathbb{1} \in [X]$), on a d'après Pythagore :

$$\underbrace{\|Y - \bar{Y}\mathbb{1}\|^2}_{SCT} = \underbrace{\|Y - \hat{Y}\|^2}_{SCR} + \underbrace{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2}_{SCE},$$

  où SCT est la "somme des carré totaux", SCR est la "somme des carrés des résidus" et SCE est la "somme des carrés expliqués".

- Dans le cas général (même si $\mathbb{1} \notin [X]$), toujours d'après Pythagore :

$$\|Y\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y}\|^2.$$

Rappel : le modèle est bon si $\hat{Y} \approx Y$ et donc si $SCR \approx 0$.

Sur la base de ces relations :

Definition (R^2, ou coeff de détermination, ou coeff de corrélation multiple)

- si $\mathbb{1} \in [X]$,
$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$
- si $\mathbb{1} \notin [X]$,
$$R^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = 1 - \frac{SCR}{\|Y\|^2}.$$

- $0 \leq R^2 \leq 1$, le modèle étant d'autant "meilleur" que R^2 est proche de 1.
- Cela n'a **aucun sens** de comparer le R^2 d'un modèle avec constante et le R^2 d'un modèle sans constante (les définitions diffèrent).
- En régression linéaire simple ($y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$), $R^2 = \hat{\rho}^2$, où $\hat{\rho}$ est la corrélation empirique entre les x_i et les y_i (cf tableau).

Le R^2 ajusté

Le R^2 a un défaut : l'ajout d'une variable $\Longrightarrow$ R^2 augmente nécessairement.
C'est parceque l'espace de projection grossit, donc forcément SCR diminue.

Pour pallier ce problème :

Definition (R_a^2 : R^2 ajusté)

- si $\mathbb{1} \in [X]$,
$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT},$$
- si $\mathbb{1} \notin [X]$,
$$R_a^2 = 1 - \frac{n}{n-p}\frac{SCR}{\|Y\|^2}.$$

- Si on ajoute une variable : $SCR \searrow$ mais $p \nearrow$ (car $p \to p+1$).
- Donc R_a^2 n'augmente que si SCR diminue fortement
- cad si la nouvelle variable ajoute un gain significatif au modèle.

Exemple : hauteur en fonction de la circonférence d'arbres

Pour le modèle de régression liant la hauteur à la circonférence des arbres, la sortie un peu plus complète du modèle sous R est :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0646     0.6517   16.98   <2e-16 ***
circ          0.2165     0.0132   16.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.132 on 98 degrees of freedom
Multiple R-squared:  0.733,      Adjusted R-squared:  0.7303
```

On y lit sur la dernière lignes :

$$R^2 = 0.733, \qquad R_a^2 = 0.7303.$$

On en déduit qu'environ 73% de la variabilité de la hauteur des arbres est expliquée par le modèle (qui ne s'appuie que sur la circonférence).

On désire tester q contraintes linéaires sur le coefficient β (vecteur de taille p). Cela s'écrit

$$H_0 : \ R\beta = 0 \quad \text{contre} \quad H_1 : \ R\beta \neq 0,$$

où R est une matrice de taille (q, p), de rang q, encodant les contraintes.

Ce test inclut, selon le choix de la matrice R (cf tableau) :

- Le *test de Student* (déjà vu) : la variable j est-elle significative ?

  $H_0 : \ \beta_j = 0$ contre $H_1 : \ \beta_j \neq 0$.

- Le *test de Fisher global* : dans un modèle avec constante ($X^{(1)} = \mathbb{1}$), y-a-t-il au moins une variable significative autre que la constante ?

  $H_0 : \ \beta_2 = \cdots = \beta_p = 0$ contre $H_1 : \exists j \in \{2, \ldots, p\}$ tel que $\beta_j \neq 0$.

- Le *Test de modèles emboités* : Le sous-modèle sans q variables (disons les q dernières, avec $q \geq 1$) est-il aussi bon que le modèle complet ?

  $H_0 : \ \beta_{p-q+1} = \cdots = \beta_p = 0$ contre H_1 : le contraire.

On note

- SCR : la SCR dans le modèle de régression sans contrainte
- SCR_c : la SCR dans le modèle de régression contraint, c'est à dire dans le sous-modèle vérifiant $R\beta = 0$.

Théorème

Si $rg(X) = p$, $rg(R) = q$ *et* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, *alors sous* $H_0 : \ R\beta = 0$ *:*

$$F = \frac{n-p}{q} \frac{SCR_c - SCR}{SCR} \sim F(q, n-p)$$

où $F(q, n-p)$ *désigne la loi de Fisher à* $(q, n-p)$ *degré de liberté.*

Preuve : Cf le tableau

D'où la région critique au niveau α pour tester $H_0 : R\beta = 0$ contre $H_1 : R\beta \neq 0$

$$RC_\alpha = \{F > f_{q,n-p}(1-\alpha)\}$$

où $f_{q,n-p}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une $F(q, n-p)$.

Test de Fisher : cas particuliers importants

Test de Student : $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Dans ce cas il y a une seule contrainte ($q = 1$).
La statistique de test est donc

$$F = (n-p)\frac{SCR_c - SCR}{SCR}$$

où

- SCR : SCR dans le modèle sans contrainte (avec toutes les variables)
- SCR_c : SCR dans le modèle contraint (sans la variable $X^{(j)}$)

On peut montrer qu'alors $F = T^2$ où $T = \hat{\beta}_j / \hat{\sigma}_{\hat{\beta}_j}$ est la statistique du test de Student de significativité de $X^{(j)}$.

Les deux tests sont donc rigoureusement équivalents dans ce cas.

Test de Fisher global : dans un modèle avec constante,

$$H_0 : \ \beta_2 = \cdots = \beta_p = 0 \text{ contre } H_1 : \exists j \in \{2, \ldots, p\} \text{ tel que } \beta_j \neq 0.$$

Il y a $q = p - 1$ contraintes dans ce cas.

Dans ce cas

- SCR : SCR dans le modèle sans contrainte (avec toutes les variables)
- SCR_c : SCR dans le modèle ne contenant que la constante.

La statistique de test peut se récrire dans ce cas comme ceci (cf tableau) :

$$F = \frac{n-p}{p-1} \frac{SCE}{SCR} = \frac{n-p}{p-1} \frac{R^2}{1-R^2}.$$

Elle ne nécessite pas de calculer SCR_c et est donc plus simple d'utilisation.

Test de Fisher : cas particuliers importants

Test de modèles emboités :

$H_0 : \beta_{p-q+1} = \cdots = \beta_p = 0$ contre H_1 : le contraire.

Il y a q contraintes.

Dans ce cas

- SCR : SCR dans le modèle sans contrainte (avec toutes les variables)
- SCR_c : SCR dans le modèle sans les q dernières variables.

Pratique du test :

- On calcule

$$F = \frac{n-p}{q} \frac{SCR_c - SCR}{SCR} \sim F(q, n-p)$$

- si $F > f_{q,n-p}(1-\alpha)$, alors on rejette H_0 au niveau α, ce qui signifie que les contraintes ne sont pas vérifiées.
- Dans ce cas, on n'accepte pas le sous-modèle par rapport au gros modèle.

Exemple : hauteur en fonction de la circonférence d'arbres

Pour le modèle de régression liant la hauteur à la circonférence des arbres, la sortie complète du modèle sous R est :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0646     0.6517   16.98   <2e-16 ***
circ          0.2165     0.0132   16.40   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.132 on 98 degrees of freedom
Multiple R-squared:  0.733,      Adjusted R-squared:  0.7303
F-statistic: 269.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

On y lit sur la dernière ligne :

- Pour le test de Fisher de significativité global : $F = 269.1$ pour $q = 1$ et $n - p = 98$ degrés de liberté, et la p-valeur du test est négligeable.
- Il y a donc au moins une variable significative (autre que la constante).
- Ici, ce test revient simplement à la significativité de `circ`.
- Il revient donc au test de Student de significativité de `circ`.
- On peut en effet vérifier que $F = T^2$ ($269.0628 = 16.40313^2$).

Pour rappel les hypothèses du modèle sont les suivantes :

- $Y = X\beta + \varepsilon$ avec
- $rg(X) = p$,
- $\mathbb{E}(\varepsilon) = 0$,
- $\mathbb{V}(\varepsilon) = \sigma^2 I_n$.

Nous allons présenter des outils de diagnostic permettant d'examiner chacune des ces hypothèses.

Nous verrons aussi des solutions envisageables lorsqu'elles sont mises en défaut.

Lien linéaire

L'hypothèse linéaire, i.e. $\mathbb{E}(Y) = X\beta$ est l'hypothèse majeure.

- Le lien linéaire entre Y et chaque variable explicative $X^{(j)}$ peut se vérifier via les nuages de points $(X^{(j)}, Y)$ et le calcul de la corrélation empirique.
- Après la modélisation, on peut visualiser les résidus $\hat{\varepsilon}$: si le lien linéaire est mis en défaut, leur comportement peut être affecté (cf plus bas).
- On peut également analyser les *résidus partiels* (non présentés ici).

Si le lien linéaire ne semble pas adapté :

- Une transformation de Y et/ou des certains variables $X^{(j)}$ peut faire apparaître un lien linéaire
- A défaut, il faut se tourner vers des modèles non linéaires.

Non-colinéarité des variables explicatives ($rg(X) = p$)

Pour rappel, $rg(X) = p$ signifie qu'aucune variable explicative $X^{(j)}$ n'est combinaison linéaire des autres.

A défaut, quel est le problème ?

- On l'a vu en début de chapitre : le paramètre β n'est alors plus défini de façon unique. Le modèle n'est donc pas identifiable.
- D'un point de vue de l'estimation : $(X'X)$ n'est pas inversible donc la formule de $\hat{\beta}$ n'a pas de sens
  (en fait il existe une infinité de $\hat{\beta}$ tel que $X'X\hat{\beta} = X'Y$)

En pratique, il est très rare qu'une variable soit parfaitement liée (linéairement) aux autres, donc en général $rg(X) = p$.

Par contre, il est possible qu'une variable soit "presque" linéairement liée aux autres... On voit tout de suite pourquoi c'est un problème.

Non-colinéarité des variables explicatives

Une variable est fortement liée aux autres si la corrélation entre cette variable est les autres est très élevée (sans être égale à 1 en valeur absolue).

Dans ce cas :

- $X'X$ est inversible, mais sa plus petite valeur propre est très proche de 0.
- Ainsi $(X'X)^{-1}$ est très instable.
- Par exemple, si on ajoute un individu (cad une ligne à X), alors $(X'X)^{-1}$ peut être radicalement différent.
- Donc $\hat{\beta} = (X'X)^{-1}X'Y$ est très instable.
- Cela est confirmé par le fait que $\mathbb{V}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, qui sera élevée.
- Il s'agit d'une situation non souhaitable d'un point de vue statistique.

Non-colinéarité des variables explicatives : le VIF

On peut détecter un éventuel problème de colinéarité en calculant les **VIF** (Variance Inflation Factor) pour chaque variable $X^{(j)}$:

1. On régresse $X^{(j)}$ par rapport aux autres variables $X^{(k)}$ $(k \neq j)$;
2. On calcule le R^2 dans cette régression, que l'on note R_j^2 ;
3. le VIF pour la variable $X^{(j)}$ vaut

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Propriétés :

- On a toujours $VIF_j \geq 1$
- Si VIF_j est élevé, cela témoigne d'un problème de multicolinéarité.
- Un seul d'alerte communément adopté est $VIF_j \geq 5$.

Sous R : fonction `vif` du package `car`

Non-colinéarité des variables explicatives : solutions

Si un problème de multicolinéarité apparait :

- On peut enlever une des variables dont le VIF est élevé.
- Quitte à choisir, autant supprimer la moins corrélée à Y.
- On peut aussi faire appel à des méthodes d'estimation robuste (régression pénalisée, non vue dans ce cours) qui évitent ce choix.

Ceci dit :

- La multicolinéarité n'est un problème que pour la qualité d'estimation de β
- Mais ce n'est pas un problème pour la prévision, car la projection $\hat{Y}$ reste bien définie et stable.

Pour rappel, le vecteur des résidus est $\hat{\varepsilon} = Y - \hat{Y} = P_{[X]^\perp}\varepsilon$.

Les résidus sont une source importante de validation du modèle.

On sait que si $Y = X\beta + \varepsilon$ avec $rg(X) = p$, $\mathbb{E}(\varepsilon) = 0$ et $\mathbb{V}(\varepsilon) = \sigma^2 I_n$, alors

- $\mathbb{E}(\hat{\varepsilon}) = 0$ et $\mathbb{V}(\hat{\varepsilon}) = \sigma^2 P_{[X]^\perp}$.
- $\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) = 0$ (cf tableau)
- et si le modèle contient une constante ($\mathbb{1} \in [X]$), alors $\bar{\hat{\varepsilon}} = 0$.

Si certaines hypothèses ne sont pas vérifiées, cela se reflète dans les résidus.

On va détailler quelques outils de diagnostics, pour :

1. évaluer graphiquement la qualité du modèle,
2. tester l'homoscédasticité (cad $\mathbb{V}(\varepsilon_i) = \sigma^2$ pour tout i),
3. tester la non-corrélation (cad $\mathbb{V}(\varepsilon)$ est une matrice diagonale),
4. examiner la normalité des résidus.

1. Nuage de points entre $\hat{Y}$ et $\hat{\varepsilon}$

Le nuage de points entre $\hat{Y}$ et $\hat{\varepsilon}$ est informatif.

Puisque $\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) = 0$, aucune structure ne devrait apparaître.

A défaut, selon les cas (cf tableau) :

- cela peut remettre en cause l'hypothèse linéaire,
- ou l'hypothèse d'homoscédasticité,
- ou l'hypothèse de non-corrélation,
- ou un mélange de tout ça...

2. Test d'homoscédasticité (cad $\mathbb{V}(\varepsilon_i) = \sigma^2$ pour tout i)

Le **test de Breusch-Pagan** permet de tester **l'homoscédasticité**.

Sous R : fonction `bptest` de la librairie `lmtest`.

Principe : on suppose que ε_i a une variance $\sigma_i^2 = \sigma^2 + z_i'\gamma$ où

- z_i est un vecteur de k variables à choisir qui pourraient expliquer l'hétéroscédasticité (si elle est présente)
- γ est un paramètre inconnu, de dimension k, à estimer.

Par défaut sous R : z_i est simplement le vecteur des variables explicatives $(X_i^{(1)}, \dots, X_i^{(p)})$, et donc $k = p$. Mais on peut proposer autre chose.

Le test de Breusch-Pagan consiste à tester

$$H_0 : \gamma = 0 \quad \text{contre} \quad H_1 : \gamma \neq 0.$$

En effet, dans ce modèle, $\gamma = 0$ ssi le bruit est homoscédastique.

(La procédure exacte du test n'est pas présentée ici)

2. Test d'homoscédasticité (cad $\mathbb{V}(\varepsilon_i) = \sigma^2$ pour tout i)

Quelles conséquences si un problème d'hétéroscédasticité apparait ?

- l'estimation par MCO de β n'est plus optimale mais reste généralement convergente.
- Les outils d'inférence (tests, ICs) deviennent faux en toute rigueur, puisqu'ils utilisent l'estimation de σ^2 (qui n'a alors pas de sens).

Que faire pour corriger le problème ?

- Une transformation de Y peut parfois remédier au problème (stabilisation de la variance par une transformation log par exemple).
- On peut parfois modéliser l'hétéroscédasticité et en tenir compte pour estimer au mieux β et pour l'inférence associée.
  Cela se fait via les MCG (moindres carrés généralisés), pas détaillés ici.

3. Non-corrélation des erreurs (cad $\mathbb{V}(\varepsilon)$ est une matrice diagonale)

La corrélation entre les ε_i peut survenir lorsque les données sont temporelles (le "i" représente le temps).

Pour le tester, on suppose que les ε_i sont "auto-corrélés" à l'ordre r, i.e.

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \cdots + \rho_r \varepsilon_{i-r} + \eta_i$$

où les η_i sont iid suivant une $\mathcal{N}(0, \sigma^2)$. Deux tests courants sont disponibles :

- **Test de Durbin-Watson** : il ne concerne que le cas $r = 1$ et teste

$$H_0 : \rho_1 = 0 \quad \text{contre} \quad H_1 : \rho_1 \neq 0.$$

- **Test de Breusch-Godfrey** : il teste (différemment),

$$H_0 : \rho_1 = \cdots = \rho_r = 0 \quad \text{contre} \quad H_1 : \text{le contraire},$$

où l'ordre r est choisi par l'utilisateur ($r = 1$ par défaut).

Dans les deux cas, H_0 **reflète l'absence d'auto-corrélation**.

Sous R : fonctions `dwtest` et `bgtest` de la librairie `lmtest`.

3. Non-corrélation des erreurs (cad $\mathbb{V}(\varepsilon)$ est une matrice diagonale)

Quelles conséquences si une auto-corrélation apparait ?

- l'estimation par MCO de β n'est plus optimale mais reste généralement convergente.
- Les outils d'inférence (tests, ICs) deviennent faux en toute rigueur.

Que faire en présence d'auto-corrélation ?

- On peut essayer de modéliser cette dépendance pour aboutir à une estimation optimale et une inférence correcte, via les MCG.
  (Mais c'est pas simple, et si on se trompe, on fait pire qu'avec les MCO)
- Ca veut surtout dire qu'on peut surement **exploiter cette dépendance** pour améliorer le modèle.
  Par exemple, expliquer Y_i à l'aide de Y_{i-1}, en plus de $X_i^{(1)}, \ldots, X_i^{(p)}$.

4. Examiner la normalité des résidus

Pour rappel :

- L'hypothèse de normalité de ε n'est pas indispensable si n est grand.
- Tous les tests restent valables asymptotiquement.
- Seul l'intervalle de prédiction exploite réellement la normalité.

Néanmoins on peut l'examiner, car si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ alors $\hat{\varepsilon} \sim \mathcal{N}(0, \sigma^2 P_{[X]^\perp})$

- On peut donc analyser la droite de Henry (qqplot) des résidus $\hat{\varepsilon}$,
- ou effectuer un test de normalité de $\hat{\varepsilon}$, comme le test de Shapiro-Wilk.

Sous R : fonctions `qqnorm` et `shapiro.test`

Analyse des individus atypiques

Un individu est atypique dans la mesure où

1) il est très mal expliqué par le modèle,
2) et/ou il influence énormément l'estimation des coefficients.

On cherche à identifier ces individus

- pour comprendre la raison de cette particularité ;
- pour éventuellement modifier le modèle en conséquences ;
- pour éventuellement exclure l'individu de l'étude.

1) Un individu i est mal expliqué par le modèle si son résidu $\hat{\varepsilon}_i$ est "anormalement" grand. Comment quantifier cet "anormalement" ?

Notons h_{ij} les éléments de la matrice $P_{[X]}$ ("hat matrix" en anglais).

Pour un modèle Gaussien, on sait que $\hat{\varepsilon}_i \sim N(0, (1-h_{ii})\sigma^2)$.

On considère donc les **résidus standardisés**

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

On s'attend à ce que $t_i \sim St(n-p)$ (c'est faux en toute rigueur, car $\hat{\varepsilon}_i \not\perp\!\!\!\perp \hat{\sigma}^2$).

Definition

On considère ainsi qu'un individu i est mal expliqué par le modèle si

$$|t_i| > t_{n-p}(1-\alpha/2)$$

pour α prédéterminé, typiquement $\alpha = 0.05$, ce qui donne $t_{n-p}(1-\alpha/2) \approx 2$.

2) Un point est influent s'il contribue beaucoup à l'estimation $\hat{\beta}$.

La valeur de h_{ii} est appelée **leverage** : elle correspond au poids de Y_i sur sa propre estimation $\hat{Y}_i$ (cf tableau).

On sait que $\sum_{i=1}^{n} h_{ii} = tr(P_{[X]}) = p$.

Donc en moyenne $h_{ii} \approx p/n$.

Definition

On dit qu'un individu i est **levier** si $h_{ii} \gg p/n$
(typiquement $h_{ii} > 2p/n$ ou $h_{ii} > 3p/n$).

La **distance de Cook** quantifie l'influence de i sur $\hat{Y}$:

$$C_i = \frac{\|\hat{Y} - \hat{Y}_{(-i)}\|^2}{p\hat{\sigma}^2}$$

où $\hat{Y}_{(-i)} = X\hat{\beta}_{(-i)}$ avec $\hat{\beta}_{(-i)}$: estimation de β sans utiliser l'individu i.

On peut montrer que

$$C_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}t_i^2.$$

Cette formule montre que la distance de Cook C_i cumule

1) l'effet "aberrant" de l'individu, au travers de la présence de t_i,
2) et son effet levier, via h_{ii}.

Sous R : fonction `cooks.distance` et dernier graphique de `plot.lm`

Exemple : consommation de glaces

On modélise la consommation de glaces aux USA, chaque semaine, en fonction du revenu moyen et de la température :

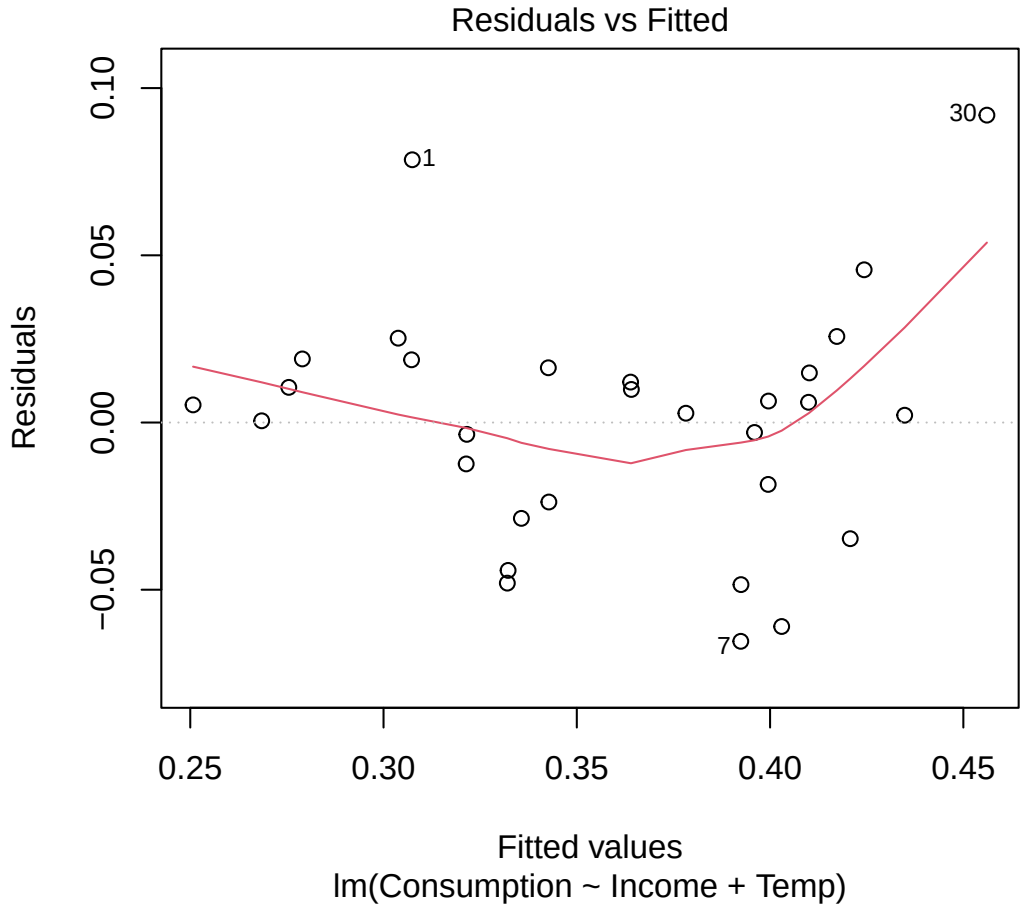
```
reg=lm(Consumption∼Income+Temp,data=tab)
```

La sortie est la suivante :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
Income       0.003530   0.001170   3.017  0.00551 **
Temp         0.003543   0.000445   7.963 1.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared:  0.7021,Adjusted R-squared:   0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

On va examiner la qualité de la modélisation.

On peut visualiser le nuage de points entre $\hat{\varepsilon}$ et $\hat{Y}$.



La répartition semble aléatoire. Aucune structure n'est apparente.

Exemple : consommation de glaces

On peut effectuer quelques vérifications numériques :

- Calcul des VIF : `vif(reg)`

```
Income      Temp
1.117863 1.117863
```

- Test d'homoscédasticité de Breusch-Pagan : `bptest(reg)`

```
BP = 2.4235, df = 2, p-value = 0.2977
```

- Test d'auto-corrélation de Durbin-Watson : : `dwtest(reg)`

```
DW = 1.0033, p-value = 0.0004485
```

- Test d'auto-corrélation de Breusch-Godfrey : `bgtest(reg)`

```
LM test = 3.7463, df = 1, p-value = 0.05292
```

Tous les résultats sont ok, sauf pour les tests d'auto-corrélation.

Une dépendance temporelle est plausible : on va essayer d'en tirer profit.

Exemple : consommation de glaces

On décide d'introduire la température de la semaine précédente (`Templag`) dans le modèle :

```
reg2=lm(Consumption~Income+Temp+Templag,data=tab)
```

On obtient :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0672394  0.0988057  -0.681  0.50243
Templag     -0.0021608  0.0007368  -2.933  0.00710 **
Income       0.0031186  0.0010429   2.990  0.00618 **
Temp         0.0053804  0.0006757   7.963 2.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03015 on 25 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8179,Adjusted R-squared:  0.796
F-statistic: 37.42 on 3 and 25 DF,  p-value: 2.131e-09
```

La modélisation semble meilleure (le R_a^2 est passé de 0.68 à 0.796).

Exemple : consommation de glaces

On peut effectuer les mêmes vérifications numériques :

- Calcul des VIF : vif(reg2)

```
 Templag   Income     Temp
4.373878 1.299288 3.893217
```

- Test d'homoscédasticité de Breusch-Pagan : `bptest(reg2)`

```
BP = 5.0063, df = 3, p-value = 0.1713
```

- Test d'auto-corrélation de Durbin-Watson : : `dwtest(reg2)`

```
DW = 1.5398, p-value = 0.03282
```

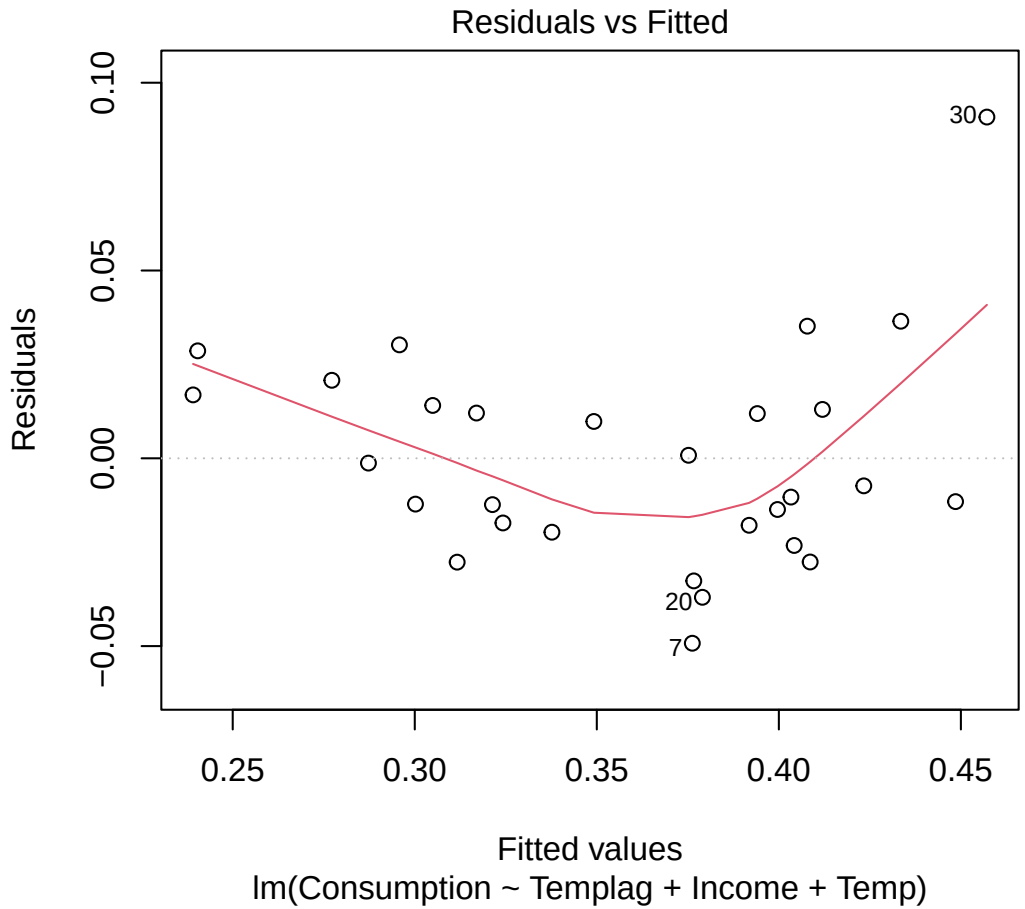
- Test d'auto-corrélation de Breusch-Godfrey : `bgtest(reg2)`

```
LM test = 0.086423, df = 1, p-value = 0.7688
```

La corrélation semble avoir été assez bien absorbée.
Les VIF ont augmenté (car Temp et Templag sont corrélés) mais pas de façon critique.

Exemple : consommation de glaces

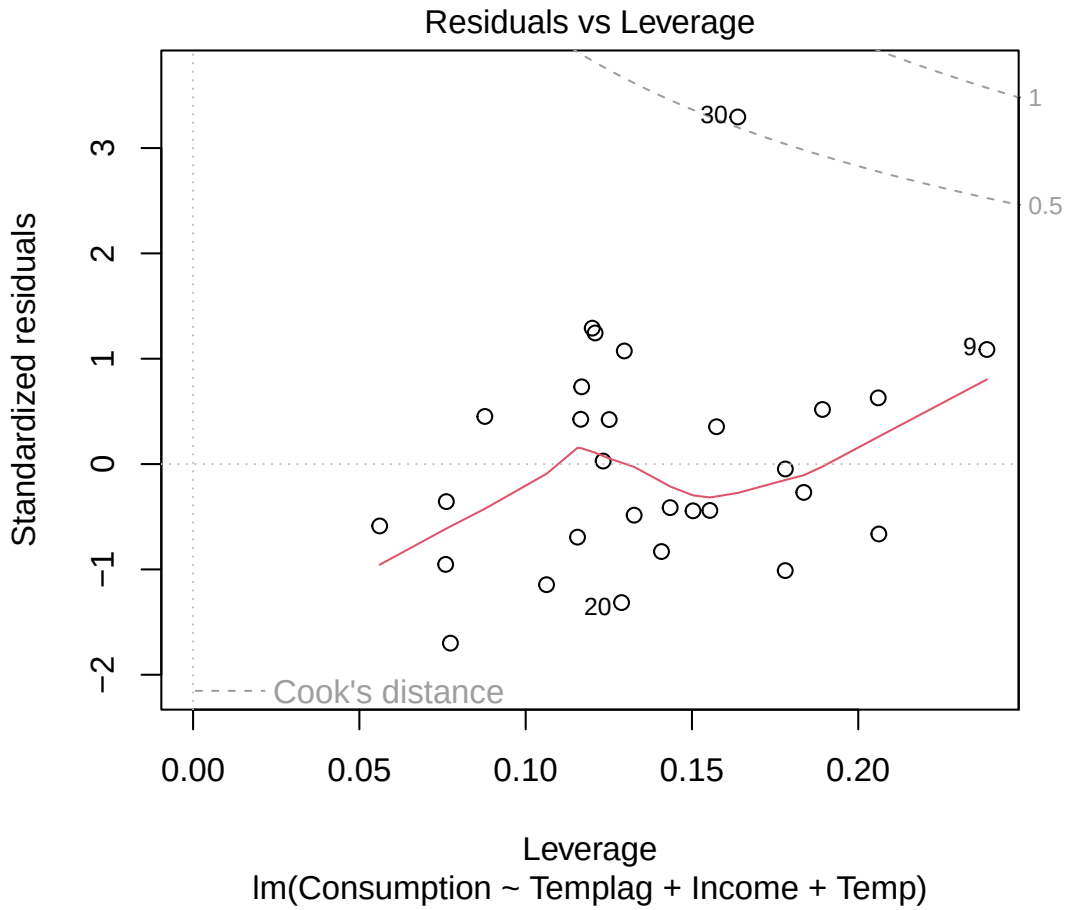
On visualise le nuage de points entre $\hat{\varepsilon}$ et $\hat{Y}$ pour ce nouveau modèle.



La répartition semble aléatoire. Aucune structure n'est apparente.
Le point 30 semble quand même un peu loin.

Exemple : consommation de glaces

Le dernier graphique de `plot(reg2)` permet d'identifier les individus atypiques.



L'individu 30 a une distance de Cook non négligeable.
Mais ce n'est pas un individu levier (son leverage est normal).
Il est juste mal expliqué par le modèle.

Exemple : consommation de glaces

Finalement, nous adoptons la modélisation suivante, pour chaque semaine i :

$$Consumption_i = -0.06 - 0.002\, Temp_{i-1} + 0.003 Income_i + 0.005\, Temp_i + \varepsilon_i$$

où ε_i représente l'erreur de modélisation, centrée et de variance $\hat{\sigma}^2 = 0.03^2$.

2 Régression linéaire

- Modélisation
- Inférence
- Validation
- Critères de sélection de modèles

En pratique, on hésite généralement entre plusieurs modèles :

- Quelles variables intégrer au modèle ?
- Comment choisir entre un modèle et un autre ?
- Idéalement : comment sélectionner le "meilleur" modèle parmi tous les sous-modèles possibles d'un gros modèle de régression linéaire ?

Plusieurs critères existent. Les principaux :

- R_a^2 : le R^2 ajusté (déjà vu)
- Test de Fisher de modèles emboités (déjà vu)
- Le C_p de Mallows
- Le critère AIC
- Le critère BIC

On les détaille ci-après.

Contexte :

- Supposons qu'on dispose de $p_{\max}$ variables explicatives, composant la matrice de design "maximale" $X_{\max}$.
- On suppose que le vrai modèle (inconnu) expliquant Y s'écrit

$$Y = X^* \beta^* + \varepsilon$$

où X^* est la sous-matrice de $X_{\max}$ formée de $p^* \leq p_{\max}$ de ses colonnes.

- On ne connait pas la valeur p^* et encore moins de quelles variables il s'agit.
- On cherche à sélectionner la bonne matrice X^*, et à estimer β^*.

En pratique, on régresse Y sur $p \leq p_{\max}$ variables, en supposant que

$$Y = X\beta + \varepsilon$$

où X : sous-matrice de $X_{\max}$ contenant les p colonnes choisies (et on obtient $\hat{\beta}$).

- Ce modèle est potentiellement faux (mauvais choix des variables).
- On cherche à calculer un score de qualité pour ce modèle.

Le R^2 ajusté (rappel)

Pour rappel, pour un modèle contenant une constante, ce critère vaut

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT},$$

où

- $SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$ ne dépend pas du modèle choisi,
- $SCR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ est propre au modèle considéré.

Entre deux modèles, on privilégie celui qui le R_a^2 **le plus élevé**.

Remarque

- Tout comme le R^2, cela n'a aucun sens de comparer le R_a^2 d'un modèle avec constante et d'un modèle sans constante (les définitions diffèrent)
- Comparer deux modèles emboités à l'aide du R^2 (au lieu du R_a^2) conduit automatiquement à choisir le plus gros : il ne faut donc pas l'utiliser.

Test de modèles emboités (rappel)

Pour choisir entre un modèle et un sous-modèle, on peut simplement appliquer le test de Fisher (cf plus haut). On calcule

$$F = \frac{n-p}{q} \frac{SCR_c - SCR}{SCR}$$

où

- SCR est le SCR du modèle le plus gros,
- SCR_c est le SCR du sous-modèle (avec moins de variables).
- p est le nombre de variables dans le gros modèle.
- q est le nombre de contraintes : le sous-modèle contient $p - q$ variables.

Si $F < f_{q,n-p}(1-\alpha)$, on privilégie le sous-modèle (H_0, au niveau α)

Si $F > f_{q,n-p}(1-\alpha)$, on privilégie le gros-modèle (H_1, au niveau α).

Remarque : Ce test permet de ne comparer que des modèles **emboités**.

Le C_p de Mallows

On rappelle qu'on suppose que le vrai modèle (inconnu) s'écrit

$$Y = X^* \beta^* + \varepsilon$$

où X^* est la sous-matrice de X_{max} formée de $p^* \leq p_{\text{max}}$ de ses colonnes.
Le modèle testé (peut-être faux) s'écrit

$$Y = X\beta + \varepsilon$$

pour lequel on obtient par MCO $\hat{\beta}$.

Le C_p de Mallows vise à estimer l'erreur de prévision

$$\mathbb{E}(\|\tilde{Y} - X\hat{\beta}\|^2)$$

pour un vecteur $\tilde{Y}$ qui suit la même loi que Y mais en est indépendant, i.e. $\tilde{Y} = X^*\beta^* + \tilde{\varepsilon}$ où $\tilde{\varepsilon}$ est indépendant de ε.

L'idée est de mesurer la performance prédictive du modèle, sur de nouveaux individus indépendants qui prendraient les mêmes valeurs pour X.

Le C_p de Mallows

Le C_p de Mallows estime l'erreur de prévision précédente comme ceci :

$$C_p = \frac{SCR}{\hat{\sigma}^2} - n + 2p$$

où

- p correspond au nombre de variables dans le modèle considéré
- SCR est le SCR du modèle considéré
- $\hat{\sigma}^2$ est l'estimation de σ^2 dans le plus gros modèle (celui contenant les $p_{\max}$ variables explicatives). Il est donc similaire quel que soit le modèle testé.

Parmi tous les modèles testés, on retient celui qui a le C_p **le plus faible**.

- Le critère AIC (Akaike Information Criterion) est motivé comme le C_p.
- Il s'intéresse de même à l'erreur de prévision $\tilde{Y} - X\hat{\beta}$.
- Mais au lieu de la quantifier via la distance quadratique, il utilise la distance de Kullback.

L'estimation de cette erreur donne :

$$AIC = n \ln \frac{SCR}{n} + 2(p+1),$$

avec les mêmes notations que ci-dessus.

Parmi tous les modèles testés, on retient celui qui a l'**AIC le plus faible**.

Remarques :

- Cette formule suppose que le modèle sous-jacent est Gaussien. C'est celle utilisée dans les logiciels.
- En pratique, l'AIC et le C_p sont très proches : ils conduisent généralement au choix du même modèle.

- Le critère BIC (Bayesian Information Criterion) est motivé différemment
- Il cherche le modèle "le plus probable" dans un formalisme Bayésien (cf le second semestre).

Mais au final, le BIC a une expression relativement proche de l'AIC :

$$BIC = n \ln \frac{SCR}{n} + (p+1) \ln n.$$

Parmi tous les modèles testés, on retient celui qui a le **BIC le plus faible**.

Remarques :

- La différence entre AIC et BIC est que le "2" devant $(p+1)$ est remplacé par $\ln n$.
- Néanmoins cette différence conduit (fréquemment) à un choix de modèle différent entre AIC et BIC (cf ci-dessous).

Lien entre les critères

Rappel :

- Un "gros" modèle a une SCR faible, mais un nombre de variables p élevé
- Un "petit" modèle a une SCR élevé, mais un nombre de variables p faible.

Tous les critères précédents essaient de trouver un compromis entre une bonne adéquation aux données (SCR faible $\Rightarrow$ modèle peu biaisé) et une petite taille du modèle (p faible $\Rightarrow$ variance d'estimation faible).

Ce **compromis biais-variance** est permanent en statistique.

En particulier, C_p, AIC et BIC consistent à minimiser une expression de la forme

$$f(SCR) + c(n)p$$

où

- f est une fonction croissante de SCR
- $c(n)p$ est un terme pénalisant les modèles ayant beaucoup de variables.

(Il existe d'autres critères construits sur le même principe.)

Lien entre les critères

En particulier $c(n) = \ln n$ pour BIC et $c(n) = 2$ pour AIC.

Dès que $\ln n > 2$ ($n \geq 8$), BIC pénalise davantage les gros modèles que AIC.

De façon générale, les critères s'ordonnent de la manière suivante en fonction de leur propension à sélectionner le modèle le plus parcimonieux :

$$BIC \leq F\ test \leq C_p \approx AIC \leq R_a^2$$

Ainsi lors d'une sélection de modèles :

- Tous les critères peuvent être d'accord sur le meilleur modèle à choisir.
- Mais s'ils diffèrent, BIC privilégiera un modèle plus petit que C_p ou AIC,
- et R_a^2 aura tendance à privilégier un modèle encore plus gros.

(Cf un exercice de TD pour une justification.)

Pour le critère BIC, lorsque $n \to \infty$:

- La probabilité qu'il sélectionne un modèle plus petit que le vrai tend vers 0 ;
- La probabilité qu'il sélectionne un modèle plus gros que le vrai tend vers 0 ;
- La probabilité qu'il sélectionne le bon modèle tend vers 1.

Pour les autres critères (C_p, AIC, R_a^2), lorsque $n \to \infty$:

- La probabilité qu'ils sélectionnent un modèle trop petit tend vers 0 ;
- La probabilité qu'ils sélectionnent un modèle trop gros **ne tend pas** vers 0 ;
- La probabilité qu'ils sélectionnent le bon modèle **ne tend pas** vers 1.

(Cf un exercice de TD pour une justification).

Sélection automatique

Etant donné $p_{\max}$ variables explicatives à disposition,

- il est tentant de tester tous les sous-modèles possibles,
- et de retenir celui qui a le BIC le plus petit (ou autre critère).

Cela représente $2^{p_{\max}}$ modèles à tester (c'est beaucoup).

Si $p_{\max}$ n'est pas trop grand, cela reste possible.

Sous R : fonction `regsubsets` de la librairie `leaps`

Attention : la sélection automatique ne garantit pas que le modèle retenu est bon. C'est simplement le meilleur modèle au sens du critère choisi. Il peut être médiocre en terme de pouvoir explicatif, comporter des problèmes de multicolinéarité, d'hétéroscédasticité ou d'auto-corrélations, par exemple.

Sélection automatique

Si $p_{\max}$ est trop grand pour se lancer dans une recherche exhaustive, on peut utiliser une procédure pas à pas (stepwise). Il en existe de plusieurs types :

Stepwise backward : (au sens d'un critère choisi, par exemple le BIC)

- on part du plus gros modèle à $p_{\max}$ variables,
- puis on enlève la variable la moins significative
- on recommence pour enlever la variable restante la moins significative
- etc
- on arrête lorsqu'aucun retrait n'améliore le modèle.

Stepwise forward : idem en partant du plus petit modèle (que la constante) et en ajoutant au fur et à mesure la variable la plus significative.

Stepwise backward hybride : comme backward, mais on essaie aussi d'ajouter une variable à chaque étape.

Stepwise forward hybride : comme forward, mais on essaie aussi d'enlever une variable à chaque étape.

Sélection automatique

- Les procédures pas à pas ne parcourent pas tous les sous-modèles possibles.
- Ils peuvent donc éventuellement "rater" le meilleur modèle
- Ils sont à utiliser si une recherche exhaustive est impossible
- La procédure la plus rapide est la forward : les petits modèles sont plus rapides à estimer.
- Les procédures hybrides sont plus lentes, mais elles parcourent davantage de modèles possibles.

Sous R : fonction `step` avec l'option `direction` égale à `backward` ou `forward` ou `both`. Par défaut le critère est AIC (option $k = 2$). On obtient BIC avec $k = \ln n$. L'option k de `step` correspond à la pénalité $c(n)$ introduite plus haut.

Exemple : consommation de glaces

On revient sur la modélisation de la consommation de glaces aux USA.

Les critères pour les deux modèles testés précédemment sont les suivants :

- `Consumption~Income+Temp`

$$BIC = -101.9 \quad AIC = -107.5 \quad C_p = 11.2 \quad R_a^2 = 0.68$$

- `Consumption~Income+Temp+Templag`

$$BIC = -108.2 \quad AIC = -115.1 \quad C_p = 4.46 \quad R_a^2 = 0.796$$

Tous les critères s'accordent à privilégier le second modèle.

Exemple : consommation de glaces

En fait, pour ce jeu de données,

- on avait accès aux 3 variables `Price`, `Income` et `Temp`.
- on y ajoute `Templag` et `Consumptionlag` pour essayer de tenir compte de la dépendance temporelle.

Une sélection automatique (exhaustive), basée sur le critère BIC, ou AIC, ou C_p, conduit au choix du modèle :

$$\texttt{Consumption} \sim \texttt{Income+Temp+Templag}$$

Le critère R_a^2 est quant à lui maximal pour le modèle :

$$\texttt{Consumption} \sim \texttt{Price+Income+Temp+Templag}$$

($R_a^2 = 0.7998$ contre $R_a^2 = 0.7960$ pour le précédent).

Etant donné la proximité des R_a^2, on pourra au final privilégier le modèle retenu par BIC, AIC et C_p, qui a l'avantage d'être plus parcimonieux.

Objectif

Dans le chapitre précédent (Régression linéaire), on a considéré que :

- la variable à expliquer Y est une variable quantitative
- les variables explicatives $X^{(j)}$ sont des variables quantitatives.

Dans ce chapitre, on suppose toujours que Y est une variable quantitative mais les variables explicatives peuvent être qualitatives et/ou quantitatives.

- Si toutes les variables explicatives $X^{(j)}$ sont qualitatives, on parle d'ANOVA (analyse de la variance)
- Si les variables explicatives mêlent à la fois des variables quantitatives et des variables qualitatives, on parle d'ANCOVA (analyse de la covariance).

On va voir que ces situations se ramènent au cas du chapitre précédent.

3. Analyse de la variance (ANOVA) et de la covariance (ANCOVA)
   - Analyse de la variance à 1 facteur
   - Analyse de la variance à 2 facteurs
   - Analyse de la variance à k facteurs
   - Analyse de la covariance (ANCOVA)

On cherche à expliquer Y à l'aide d'une seule variable qualitative A.

Notations :

- Y : variable quantitative observée auprès de n individus
- A : variable explicative qualitative (on dit aussi "facteur") composée de I modalités notées $A_1, \ldots, A_I$, observée sur les mêmes individus.
- n_i : effectif dans la modalité A_i, $i = 1, \ldots, I$.
- $Y_{i,j}$: valeur de Y pour l'individu j appartenant à la modalité A_i, pour $i = 1, \ldots, I$ et $j = 1, \ldots, n_i$.
- Y_k : valeur de Y pour l'individu k, $k = 1, \ldots, n$ (cette notation ne tient pas compte de l'appartenance de l'individu à sa modalité pour A).
- μ_i : espérance de Y dans la classe A_i, i.e. $\mu_i = \mathbb{E}(Y|A_i)$.

Question :
Le facteur A a-t-il une influence sur Y ? Plus précisément a-t-on $\mu_1 = \cdots = \mu_I$?

Le modèle de base s'écrit, pour tout $i = 1, \dots, I$ et $j = 1, \dots, n_i$:

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}$$

où $\varepsilon_{i,j}$ sont des variables centrées, non-corrélées 2 à 2, et de variance σ^2. Ainsi

- Y est aléatoire et admet comme espérance μ_i dans la modalité A_i
- La variance de Y est similaire quelle que soit la modalité A_i.
- On va chercher à savoir si les μ_i sont différents.

En utilisant la notation Y_k au lieu de $Y_{i,j}$, ce modèle s'écrit également :

$$Y_k = \sum_{i=1}^{I} \mu_i \mathbb{1}_{A_i}(k) + \varepsilon_k, \quad k = 1, \dots, n,$$

où les ε_k sont centrées, non-corrélées 2 à 2, et de variance σ^2.

De façon matricielle, le modèle s'écrit donc

$$Y = X\mu + \varepsilon$$

où X matrice de taille (n, I), μ de taille I, et ε de taille n, sont définis par

$$X = \left(\mathbb{1}_{A_1} | \dots | \mathbb{1}_{A_I}\right), \qquad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Il s'agit d'un modèle de régression linéaire standard dans lequel :

- toutes les variables sont quantitatives et ne prennent que les valeurs 0 ou 1,
- il n'y a pas de constante.

Le modèle général contenant une constante m s'écrirait

$$Y_k = m + \sum_{i=1}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \varepsilon_k, \qquad k = 1, \dots, n,$$

ou de façon matricielle : $Y = \tilde{X}\beta + \varepsilon$ avec

$$\tilde{X} = \left(\mathbb{1} | \mathbb{1}_{A_1} | \dots | \mathbb{1}_{A_I}\right), \qquad \beta = \begin{pmatrix} m \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix}.$$

- Les coefficients α_i n'ont aucune raison de coincider avec $\mu_i = \mathbb{E}(Y|A_i)$.
- Ce modèle n'est pas de plein rang car $\mathbb{1} = \sum_{i=1}^{I} \mathbb{1}_{A_i}$.
- Il faut donc ajouter une contrainte sur β pour le rendre identifiable.

Exemples de contraintes dans le modèle général précédent :

1) $m = 0$:

- On retrouve alors le modèle initial sans constante.
- Dans ce cas $\alpha_i = \mu_i = \mathbb{E}(Y|A_i)$.
- Sous R : on impose cette contrainte avec la commande `lm(Y~A-1)`.

2) $\alpha_1 = 0$:

- dans ce cas, l'interprétation des coefficients est différente :

$$m = \mu_1 \quad \text{et} \quad \alpha_i = \mu_i - \mu_1 \text{ pour tout } i = 2, \ldots, I.$$

- Sous R : il s'agit de la contrainte par défaut dans `lm(Y~A)`.

Moralité : en présence d'un facteur, l'interprétation des coefficients est différente selon la contrainte choisie.

Mais quelle que soit la contrainte choisie, l'estimation de l'espérance de Y dans chaque classe A_i est identique (et naturelle). Idem pour la variance.

Proposition

Dans le modèle général précédent de l'ANOVA à 1 facteur ($Y = \tilde{X}\beta + \varepsilon$), quelle que soit la contrainte linéaire choisie, l'estimation par MCO des paramètres conduit à

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} = \bar{Y}_i, \quad i = 1, \dots, I$$

et l'estimateur sans biais de la variance est

$$\hat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2.$$

Preuve : Cf le tableau

Pour rappel, on souhaite tester $H_0 : \mu_1 = \cdots = \mu_I$.

C'est un test de contraintes linéaires dans le modèle de régression précédent.
On retrouve le test d'analyse de la variance présenté en introduction :

Proposition

Si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, *alors sous* $H_0 : \mu_1 = \cdots = \mu_I$,

$$F = \frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)} \sim F(I-1, n-I)$$

où $S^2_{inter} = \sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2$ *et* $S^2_{intra} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_i)^2$.
D'où la région critique au niveau α : $RC_\alpha = \{F > f_{I-1,n-I}(1-\alpha)\}$.

Preuve : Il y a $I-1$ contraintes à tester. La statistique de test s'écrit donc

$$F = \frac{n-I}{I-1}\frac{SCR_c - SCR}{SCR}.$$

On montre que $SCR = S^2_{intra}$, $SCR_c = S^2_T = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y})^2$ et on utilise la relation $S^2_T = S^2_{inter} + S^2_{intra}$.

Sous R : `anova(lm(Y~A))`

Significativité du facteur : test d'analyse de la variance

- Le test d'analyse de la variance précédent teste l'égalité des **espérances** entre modalité (et non celle des variances).
- Il est valable sous l'hypothèse $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
- L'hypothèse Gaussienne n'est pas critique si n est grand
- Par contre l'homoscédasticité est importante
- On peut l'examiner en testant l'égalité des variances dans chaque modalité
- Cela se fait avec le test de Levene ou le test de Bartlett
- Sous R : `leveneTest` ou `bartlett.test` de la librairie `car`

Analyse post-hoc : problème des tests multiples

Si le facteur A est significatif, on souhaite en savoir davantage : quelle(s) modalité(s) diffère(nt) des autres ? Pour cela, on désire effectuer tous les tests

$$H_0^{i,j} : \mu_i = \mu_j \quad \text{versus} \quad H_1^{i,j} : \mu_i \neq \mu_j$$

pour tout $i \neq j$ dans $\{1, \ldots, I\}$, ce qui correspond à $I(I-1)/2$ tests.

- Une idée naive consisterait à effectuer tous les tests de Student de comparaison des moyennes, chacun au niveau α.
- Mais vu le nombre de tests, cela conduirait à de nombreux **faux positifs**. (Explications dans le poly et au tableau)
- Les faux positifs sont un problème bien connu des tests multiples.
- Il faut apporter une correction à la règle de décision.
- Des corrections "universelles" existent : correction de Bonferroni (cf poly et tableau), correction de Benjamin Hochberg, ...
- Pour l'ANOVA à 1 facteur, le **test de Tukey** répond au problème.

La statistique de Tukey est

$$Q = \sqrt{2} \max_{(i,j)} \frac{|\bar{Y}_i - \bar{Y}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

Sous $H_0 : \mu_1 = \cdots = \mu_I$ et en supposant $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$,

$$Q \sim Q_{I,n-I}$$

où $Q_{I,n-I}$ désigne la loi de Tukey à $(I, n-I)$ degrés de liberté.
(Pour être précis : ceci est exact si tous les n_i sont égaux, sinon la loi est approximativement la loi de Tukey)

Pour tester chaque $H_0^{i,j} : \mu_i = \mu_j$, on utilise alors les régions critiques

$$RC_\alpha^{i,j} = \left\{ |\bar{Y}_i - \bar{Y}_j| > \frac{\hat{\sigma}}{\sqrt{2}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, Q_{I,n-I}(1-\alpha) \right\},$$

où $Q_{I,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi $Q_{I,n-I}$.

La forme des $RC_\alpha^{i,j}$ précédents assure que (cf poly et tableau)

$$\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) = \alpha.$$

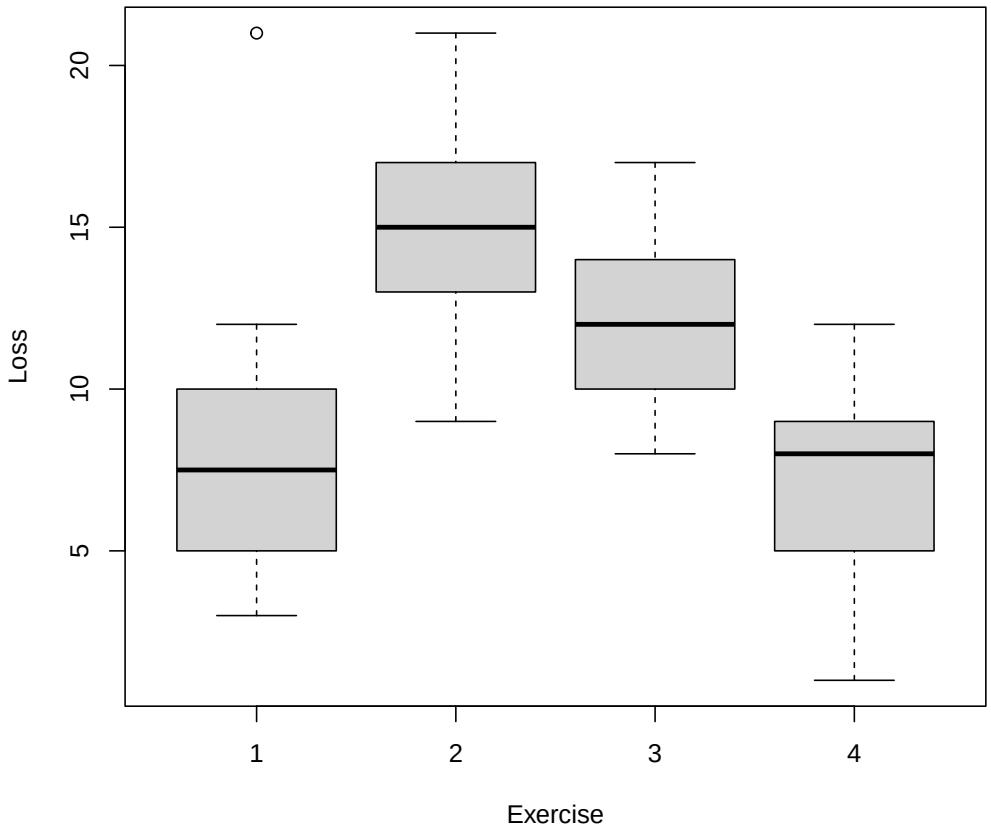
- Ainsi : si toutes les hypothèses nulles $H_0^{i,j}$ sont vérifiées ($\mu_1 = \cdots = \mu_I$), alors la probabilité d'observer des faux positifs (ne serait-ce qu'un) vaut α.
- Cela revient à la probabilité de conclure au moins une hypothèse $H_1^{i,j}$, soit $\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right)$, en accord avec la formule précédente.
- On dit que le niveau **simultané** de première espèce vaut α
- En comparaison, chaque test de Student au niveau α pour tester $H_0^{i,j}$ garantit uniquement $\mathbb{P}_{\mu_i=\mu_j}(H_1^{i,j}) = \alpha$, mais si on les cumule on aura $\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) \approx 1$, autrement dit l'apparition de faux positifs.
- Avec le test de Tukey, deux moyennes significativement différentes le sont vraiment, et non pas uniquement à cause de l'apparition de faux positifs.

Sous R : fonction `TukeyHSD`

Exemple : perte de poids

On cherche à expliquer la perte de poids (`Loss`) en fonction du programme d'activité physique choisi (`Exercise` : 4 possibilités).

Voici les boxplots de `Loss` par modalité de `Exercise`



Exemple : perte de poids

On souhaite tester l'égalité des moyennes par une ANOVA.

- On commence par vérifier l'égalité des variances dans chaque modalité :

```
leveneTest(Loss~Exercise)
```

On obtient :

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.6527  0.584
      68
```

Il n'y a pas de raison de rejeter l'hypothèse d'égalité des variances.

- On effectue à présent l'ANOVA :

```
reg=lm(Loss~Exercise)
anova(reg)
```

On obtient :

```
Response: Loss
          Df Sum Sq Mean Sq F value    Pr(>F)
Exercise   3 712.56 237.519  20.657 1.269e-09 ***
Residuals 68 781.89  11.498
```

Exemple : perte de poids

Les moyennes sont significativement différentes. On lit en particulier :

$$I-1=3,\ n-I=68,\ S^2_{inter}=712.56,\ S^2_{intra}=781.89,\ F=\frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)}=20.657$$

On termine en analysant plus finement les différences de moyennes :

```
TukeyHSD(aov(Loss~Exercise))
```

On obtient :

```
$Exercise
          diff         lwr        upr     p adj
2-1  7.1666667   4.1897551 10.1435782 0.0000001
3-1  3.8888889   0.9119773  6.8658005 0.0053823
4-1 -0.6111111  -3.5880227  2.3658005 0.9487355
3-2 -3.2777778  -6.2546894 -0.3008662 0.0252761
4-2 -7.7777778 -10.7546894 -4.8008662 0.0000000
4-3 -4.5000000  -7.4769116 -1.5230884 0.0009537
```

Toutes les différences sont significatives, sauf entre l'exercice 4 et l'exercice 1.

3. Analyse de la variance (ANOVA) et de la covariance (ANCOVA)
   - Analyse de la variance à 1 facteur
   - Analyse de la variance à 2 facteurs
   - Analyse de la variance à k facteurs
   - Analyse de la covariance (ANCOVA)

On cherche à expliquer Y à l'aide de 2 variables qualitatives A et B.

Notations :

- Y : variable quantitative observée auprès de n individus
- A : facteur composé de I modalités notées $A_1, \ldots, A_I$.
- B : facteur composé de J modalités notées $B_1, \ldots, B_J$.
- n_{ij} : effectif dans la modalité $A_i \cap B_j$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.
- Y_{ijk} : valeur de Y pour l'individu k appartenant à la modalité A_i et à la modalité B_j, pour $i = 1, \ldots, I$, $j = 1, \ldots, J$ et $k = 1, \ldots, n_{ij}$.
- Y_k : valeur de Y pour l'individu k, $k = 1, \ldots, n$ (cette notation ne tient pas compte de l'appartenance de l'individu aux modalités de A et B).
- μ_{ij} : espérance de Y dans la classe $A_i \cap B_j$, i.e. $\mu_{ij} = \mathbb{E}(Y | A_i \cap B_j)$.
- $\mu_{i.}$ et $\mu_{.j}$: espérances marginales dans A_i et B_j, respectivement.

Modélisation pour l'ANOVA à 2 facteurs

Le modèle général liant Y aux 2 facteurs A et B s'écrit :

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

pour $i = 1, \ldots, I$, $j = 1, \ldots, J$ et $k = 1, \ldots, n_{ij}$, où les ε_{ijk} sont des variables centrées, non-corrélées 2 à 2, et de variance σ^2.

On décompose l'espérance μ_{ij} dans $A_i \cap B_j$, de façon additive en :

- m : l'effet moyen de Y (sans tenir compte de A et B),
- $\alpha_i = \mu_{i.} - m$: l'effet marginal dû à A,
- $\beta_j = \mu_{.j} - m$: l'effet marginal dû à B,
- $\gamma_{ij} = \mu_{ij} - m - \alpha_i - \beta_j$: l'effet restant, dû à l'interaction entre A et B.

Avec ces notations, le modèle s'écrit :

$$Y_{ijk} = m + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$\longrightarrow$ On cherche à évaluer la significativité de chaque effet.

Exemple 1 : Y : satisfaction des employés ; A : type d'horaires (flexibles ou fixes) ; B : niveau de formation (base ou avancé). On peut imaginer :

- un effet dû à A : la satisfaction est plus grande en horaires flexibles,
- un effet dû à B : la satisfaction est plus élevée pour une formation avancée,
- pas d'interaction particulière entre A et B.

Exemple 2 : Y : rendement d'une plante ; A : type d'engrais (1 ou 2) ; B : quantité d'eau (faible, moyenne, élevée). On peut imaginer :

- Un effet dû à A : le rendement diffère selon l'engrais utilisé.
- Un effet dû à B : le rendement est meilleur lorsqu'il y a beaucoup d'eau.
- Une interaction : l'engrais 1 est meilleur avec peu d'eau, et inversement pour l'engrais 2.

On peut aussi imaginer que l'interaction est tellement forte que l'effet dû à A semble absent : les deux engrais semblent avoir le même effet en moyenne.

On peut récrire le modèle sous la forme plus standard d'un modèle de régression linéaire avec des variables indicatrices : pour tout $k = 1, \dots, n$,

$$Y_k = m + \sum_{i=1}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \sum_{j=1}^{J} \beta_j \mathbb{1}_{B_j}(k) + \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij} \mathbb{1}_{A_i \cap B_j}(k) + \varepsilon_k$$

- Il y a plusieurs problèmes de multicolinéarité dans ce modèle.
- Le problème initial de l'ANOVA à 2 facteurs fait intervenir $I \times J$ inconnues (les espérances μ_{ij} de Y dans les $A_i \cap B_j$).
- Or le modèle précédent contient $1 + I + J + IJ$ paramètres.
- Il faut donc $1 + I + J$ contraintes pour rendre le modèle identifiable.
- L'interprétation précédente des coefficients m, α_i, β_j et γ_{ij} en fonction de μ_{ij}, $\mu_{i.}$ et $\mu_{.j}$, est liée aux contraintes :

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \forall j = 1, \dots, J, \sum_{i=1}^{I} \gamma_{ij} = 0, \quad \forall i = 1, \dots, I, \sum_{j=1}^{J} \gamma_{ij} = 0.$$

Sous R :

- Le modèle complet (avec interaction) se lance avec la commande

  `lm(Y~A+B+A:B)`

  ou de façon équivalente

  `lm(Y~A*B)`

- Par défaut les contraintes ne sont pas celles (naturelles) précédentes, mais

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad \forall j = 1, \dots, J, \ \gamma_{1j} = 0, \quad \forall i = 1, \dots, I, \ \gamma_{i1} = 0.$$

- Avec ces contraintes, l'interprétation des paramètres est la suivante :

$$m = \mu_{11}, \quad \alpha_i = \mu_{i1} - \mu_{11}, \quad \beta_j = \mu_{1j} - \mu_{11}, \quad \gamma_{ij} = \mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11}.$$

- Il faut donc être très vigilant dans l'interprétation des coefficients : tout dépend des contraintes choisies.
- Pour imposer les contraintes du slide précédent :

  `lm(Y~A*B,contrasts=(A=contr.sum,B=contr.sum))`

Comme pour l'ANOVA à 1 facteur, le choix des contraintes n'affecte pas l'estimation de l'espérance de Y dans chaque modalité croisée $A_i \cap B_j$.

Proposition

Dans le modèle précédent, quelle que soient les contraintes linéaires choisies, l'estimation par MCO conduit à l'estimation, pour tout $i = 1, \dots, I$, $j = 1, \dots, J$ *et* $k = 1, \dots, n_{ij}$,

$$\hat{Y}_{ijk} = \bar{Y}_{ij}$$

où $\bar{Y}_{ij}$ *désigne la moyenne empirique dans la modalité croisée* $A_i \cap B_j$ $(\bar{Y}_{ij} = n_{ij}^{-1} \sum_{k=1}^{n_{ij}} Y_{ijk})$, *et à l'estimation de la variance résiduelle :*

$$\hat{\sigma}^2 = \frac{1}{n - IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})^2.$$

Tests de significativité des effets

Dans une ANOVA à 2 facteurs, on souhaite tester :

- si l'effet dû à l'interaction entre A et B est significatif,
- si l'effet marginal dû à A est significatif,
- si l'effet marginal dû à B est significatif.

Pour cela on part du modèle complet (avec interaction) et on commence par tester la significativité de l'interaction :

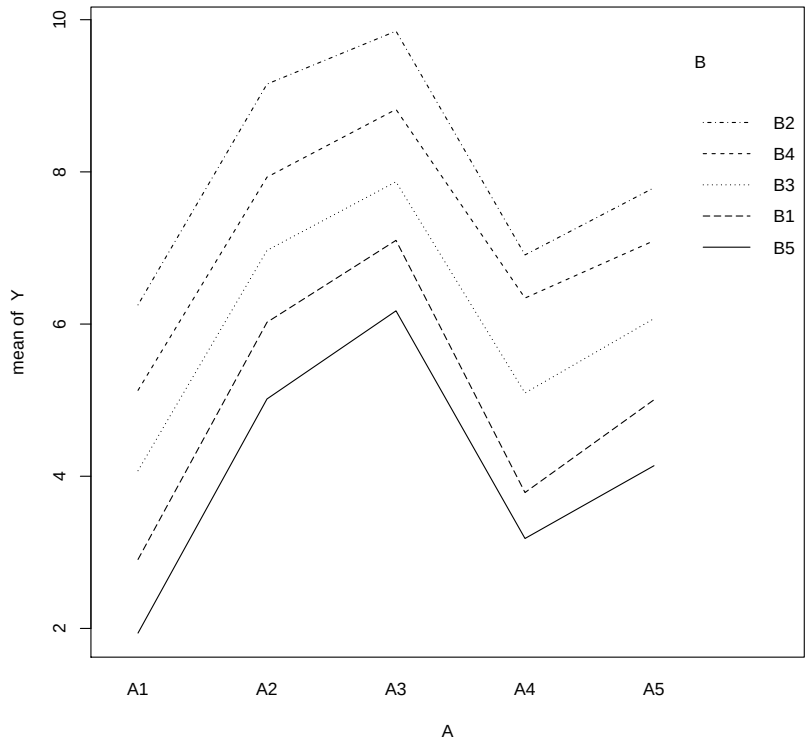
- A-t-on $\gamma_{ij} = 0$ pour tout i, j ?
- Dans ce dernier cas, on dit que le modèle est **additif** car alors

$$Y_{ijk} = m + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

Examen graphique de la présence d'interaction

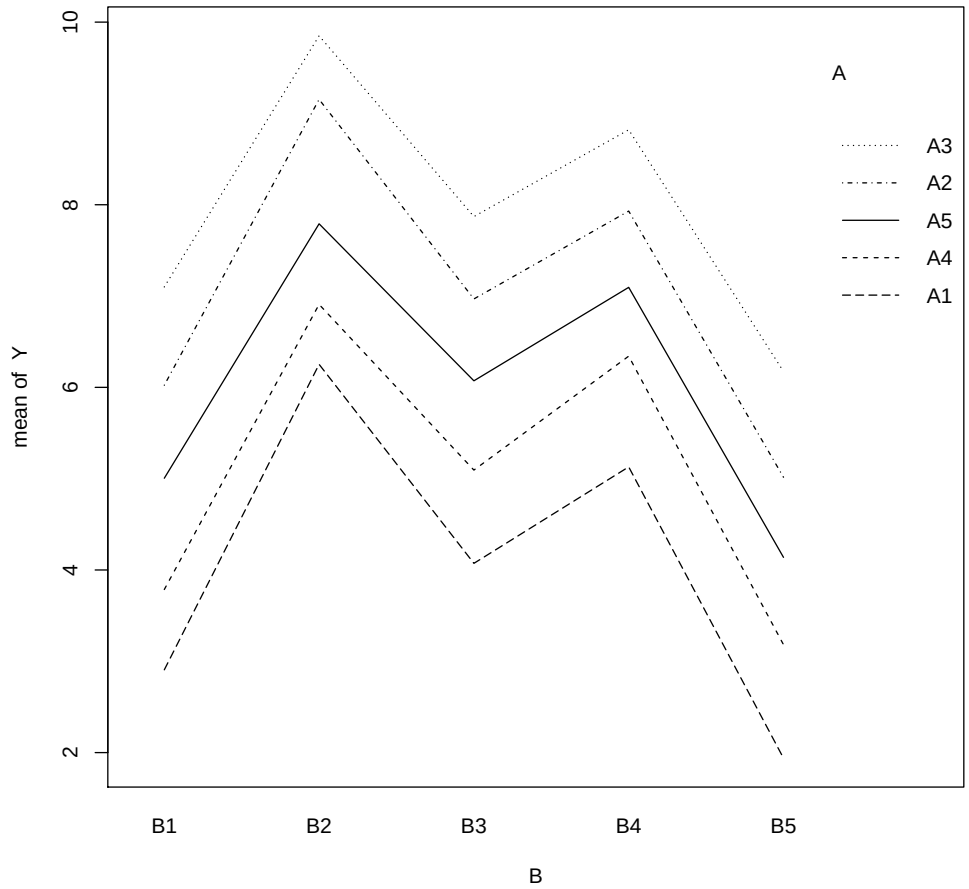
On peut représenter les "interaction plot" :

- Il s'agit de la moyenne de Y selon les modalités croisées $A_i \cap B_j$
- On place en abscisse un facteur (A) et en ordonnée l'autre facteur (B).
- Et on relie les moyennes par modalité de B.
- **S'il n'y a pas d'interaction, les courbes sont plus ou moins parallèles.** (Explication dans le poly et au tableau)



Examen graphique de la présence d'interaction

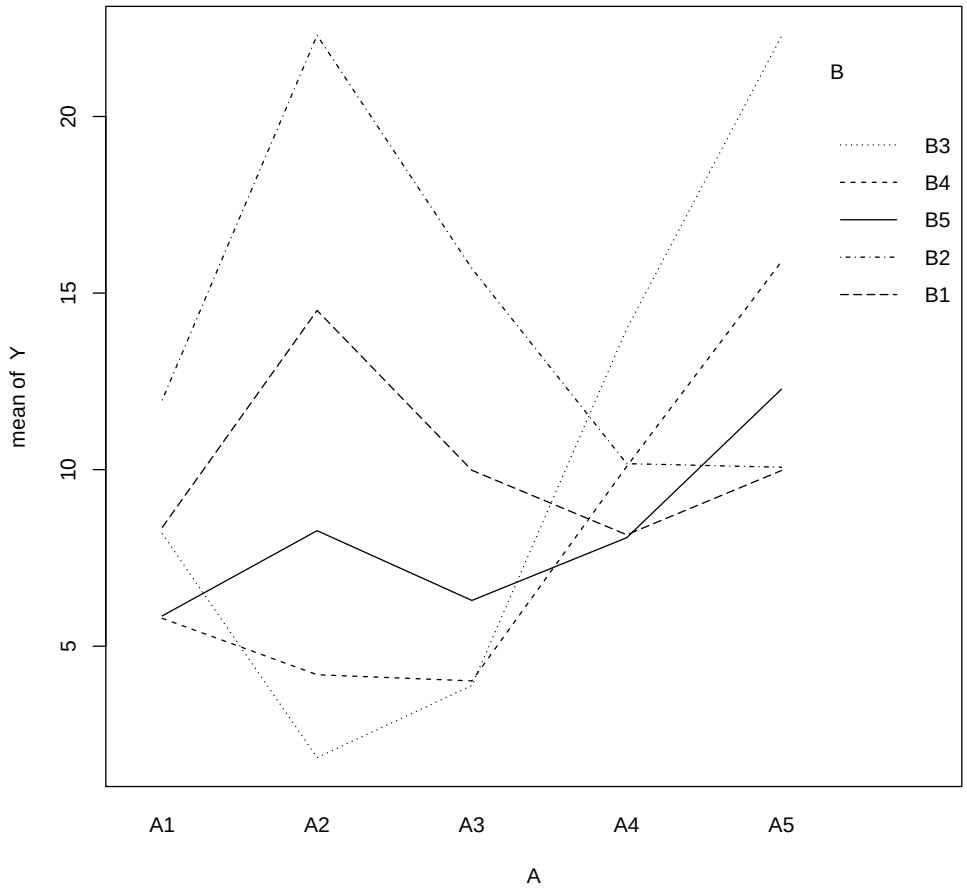
On peut évidemment inverser le rôle joué par A et B



Sous R : `interaction.plot(B,A,Y)` met le facteur B en abscisse.

Examen graphique de la présence d'interaction

Exemple de situation où il y a une interaction :



- On désire tester la présence d'une interaction :

$$H_0^{(AB)} : \gamma_{ij} = 0 \text{ pour tout } i, j$$

- Si on conclut $H_0^{(AB)}$, on désire alors tester les effets marginaux

$$H_0^{(A)} : \alpha_i = 0 \text{ pour tout } i \quad \text{et} \quad H_0^{(B)} : \beta_j = 0 \text{ pour tout } j,$$

- Si on rejette $H_0^{(AB)}$, cela n'a aucun sens de tester si A ou B ont un effet : ils en ont un au travers de leur interaction.

Ces tests se ramènent à un test de contraintes dans le modèle de régression.

On suppose que "le plan est équilibré" : cela signifie que n_{ij} est identique quel que soit i et j. (A défaut, tout se complique).

Dans ce cas, on a la formule d'analyse de la variance :

$$S_T^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

où

- $S_T^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y})^2$ est la SCT,
- $S_A^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (\bar{Y}_{i.} - \bar{Y})^2$ est l'équivalent de S_{inter}^2 dans le cas de l'ANOVA à 1 facteur dont le facteur est A,
- $S_B^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (\bar{Y}_{.j} - \bar{Y})^2$ est l'équivalent de S_{inter}^2 dans le cas de l'ANOVA à 1 facteur dont le facteur est B,
- $S_{AB}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y})^2$ quantifie l'interaction,
- $S_R^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})^2$ est l'équivalent de S_{intra}^2 dans l'ANOVA à 1 facteur.

Si le plan est équilibré, la statistique $F = \frac{n-p}{q}(SCR_c - SCR)/SCR$ dans le test de contraintes linéaires se simplifie en :

- pour tester l'interaction, $H_0^{(AB)} : \gamma_{ij} = 0$ pour tout i, j,

$$F^{(AB)} = \frac{S_{AB}^2/(I-1)(J-1)}{S_R^2/(n-IJ)}$$

- pour tester l'effet marginal de A, $H_0^{(A)} : \alpha_i = 0$ pour tout i,

$$F^{(A)} = \frac{S_A^2/(I-1)}{S_R^2/(n-IJ)}$$

- pour tester l'effet marginal de B, $H_0^{(B)} : \beta_j = 0$ pour tout j,

$$F^{(B)} = \frac{S_B^2/(J-1)}{S_R^2/(n-IJ)}$$

Si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, ces statistiques suivent respectivement sous l'hypothèse nulle la loi de Fisher $F((I-1)(J-1), n-IJ)$, $F(I-1, n-IJ)$ et $F(J-1, n-IJ)$.

Dans les logiciels, ces tests sont résumés dans un tableau comme ci-dessous.

Sous R : `anova(lm(Y~ A*B))`

	dll	SC	mean SC	F	p-value
A	$I-1$	S_A^2	$S_A^2/(I-1)$	$F^{(A)}$	...
B	$J-1$	S_B^2	$S_B^2/(J-1)$	$F^{(B)}$	...
AB	$(I-1)(J-1)$	S_{AB}^2	$S_{AB}^2/(I-1)(J-1)$	$F^{(AB)}$	...
Résidus	$n-IJ$	S_R^2	$S_R^2/(n-IJ)$		

- Les tests de Fisher reposent sur l'hypothèse $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
- La normalité n'est pas critique, mais l'homoscédasticité oui
- On peut tester l'égalité des variances dans chaque modalité de A (ou de B), ou dans chaque modalité croisée si les n_{ij} sont suffisamment grands.
- Cela peut se faire avec le test de Levene ou le test de Bartlett.

En pratique :

- On teste l'égalité des variances.
- On forme le tableau précédent (indépendant des contraintes choisies).
- Si l'interaction AB est significative, on ne touche à rien.
- Si elle n'est pas significative, on analyse les effets marginaux de A et B.
- S'ils sont significatifs, le modèle est additif et revient à `lm(Y~A+B)`.
- Sinon, on peut enlever A (ou B) du modèle.
- Une fois les effets identifiés, on peut faire une analyse post-hoc en examinant plus finement les différences entre modalités (croisées), via le test de Tukey comme en ANOVA à 1 facteur.

Exemple : perte de poids

On reprend l'exemple de la perte de poids (`Loss`). Deux facteurs sont en réalité observés :

- le programme d'activité physique choisi (4 possibilités),
- le régime alimentaire suivi (3 possibilités).

Les détails de l'étude sont disponibles dans Moodle. On conclut :

- Le programme d'activité physique a un effet significatif ;
- Le régime alimentaire : pas trop en moyenne ;
- Il y a une forte interaction : certaines activités physiques sont très efficaces lorsqu'elles sont couplées à un certain régime alimentaire, mais pas du tout avec d'autres.

3. Analyse de la variance (ANOVA) et de la covariance (ANCOVA)
   - Analyse de la variance à 1 facteur
   - Analyse de la variance à 2 facteurs
   - Analyse de la variance à k facteurs
   - Analyse de la covariance (ANCOVA)

Principe et limites

On cherche à expliquer Y à l'aide de k variables qualitatives $A, B, C, \ldots$

- On peut supposer que l'espérance de Y dépend de chaque facteur,
- et de leur interaction 2 à 2 : AB, BC, AC, etc,
- mais aussi de leur interaction triple : ABC, etc,
- voire davantage,
- cela fait $2^k - 1$ effets possibles pour k facteurs.
- On peut tester chaque effet avec un test de contraintes linéaires.
- Mais cela fait beaucoup de tests.
- Et l'effectif dans chaque modalité croisée risque d'être trop faible.
- En pratique : on fait des choix pour ne faire intervenir qu'un nombre limité d'effets et d'interactions.

3. Analyse de la variance (ANOVA) et de la covariance (ANCOVA)
   - Analyse de la variance à 1 facteur
   - Analyse de la variance à 2 facteurs
   - Analyse de la variance à k facteurs
   - Analyse de la covariance (ANCOVA)

Il s'agit de la situation la plus générale : on cherche à expliquer Y à l'aide de variables quantitatives et qualitatives.

Le modèle pourra ainsi inclure :

- les effets de chaque variable quantitative (via chaque coefficient de régression β_j associé),
- les effets des facteurs et des interactions entre les facteurs (comme pour l'ANOVA),
- mais aussi les effets des interactions entre les facteurs et les variables quantitatives.
- Par exemple on peut imaginer que le coefficient β_j de la variable quantitative $X^{(j)}$ prend en réalité deux valeurs différentes selon qu'on est dans la première modalité du facteur A ou dans la seconde : il s'agit d'une interaction entre $X^{(j)}$ et A.
- On peut tester chaque effet par un test de Fisher
- Evidemment, il faut faire des choix...

Exemple : régression simple avec un facteur

Sous R : Si Y est la variable réponse quantitative, X une variable quantitative et A un facteur à I modalités,

- `lm(Y~X+A)` estime le modèle sans interaction :

$$Y_k = m + \beta X_k + \sum_{i=2}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \varepsilon_k$$

pour chaque individu $k = 1, \ldots, n$.
(La contrainte $\alpha_1 = 0$ est adoptée pour rendre le modèle identifiable.)

- `lm(Y~X+A+X:A)` ou `lm(Y~X*A)` estime le modèle avec interaction :

$$Y_k = m + \beta X_k + \sum_{i=2}^{I} \beta_i X_k \mathbb{1}_{A_i}(k) + \sum_{i=2}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \varepsilon_k.$$

Ici le coefficient β devant X se décline différemment selon la modalité A_i.

$\longrightarrow$ *Une activité de présentation de l'ANCOVA et de mise en oeuvre sous R (via une vidéo) est disponible dans Moodle.*

4 Modèles linéaires généralisés

- Généralité sur les GLM (generalized linear models)
  - Limites du modèle linéaire
  - Vers le modèle linéaire généralisé : 3 cas fondamentaux
  - Le modèle linéaire généralisé
- Le modèle logistique pour Y binaire
- Modèles pour données catégorielles
- Modèles pour données de comptage

Rappel : objectif de la régression

Y : variable aléatoire
$X = (X^{(1)}, \ldots, X^{(p)})$: vecteur de p variables aléatoires

But : expliquer/approcher/prédire au mieux Y en fonction de X.

Mathématiquement : on cherche la "meilleure" fonction de X qui approche Y.
La solution (dans L^2) est connue, il s'agit de : $\mathbb{E}(Y|X)$.

Statistiquement : on cherche à estimer $\mathbb{E}(Y|X = x)$, pour tous les x pertinents, à partir de n réalisations du couple (Y, X).

Mais estimer une fonction de p variables est un peu ambitieux.

$\Longrightarrow$ on fait généralement des hypothèses sur la forme de $\mathbb{E}(Y|X = x)$.

En régression **linéaire** on suppose que pour un certain paramètre $\beta \in \mathbb{R}^p$:

$$\mathbb{E}(Y|X) = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} = X'\beta.$$

Remarques sur les notations :

- Ici $X = (X^{(1)}, \ldots, X^{(p)})$ désigne le vecteur des variables explicatives, et non la matrice de design observée.
- Nous considérons ici, et dans la suite, que X est aléatoire, ce qui est le cas le plus général.
- Lorsque cela sera utile, nous noterons $\mathbf{X}$ la matrice de design, qui contient en ligne les n réalisations de X, et $\mathbf{Y}$ le vecteur des n réalisations de Y.
- En particulier, l'estimateur par MCO de β dans le modèle linéaire, basée sur n réalisations s'écrit donc avec ces notations $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

L'hypothèse $\mathbb{E}(Y|X) = X'\beta$ du modèle de régression linéaire implique que $\mathbb{E}(Y|X)$ peut prendre **n'importe quelle valeur réelle**.

- Ce n'est pas une restriction si $Y|X$ suit une loi Gaussienne, ou tout autre loi continue sur $\mathbb{R}$;
- **Cela n'est pas adpaté pour certaines variables** Y, notamment si Y est qualitative ou discrète.

Par exemple, l'hypothèse linéaire est inadaptée si Y représente :

- le retour à l'emploi en moins de 3 mois ($Y = 0$ ou 1) ;
- l'efficacité d'un traitement médical ($Y = 0$ ou 1) ;
- la catégorie d'un client sollicitant un prêt bancaire ($Y = 0$ ou 1).
- le segmentation d'une clientèle en k catégories ($Y \in \{A_1, \ldots, A_k\}$) ;
- le nombre de morts par accidentologie routière en 1 mois ($Y \in \mathbb{N}$).

Objectif d'un modèle GLM

Dans ces exemples, le but est toujours de lier Y à $X = (X^{(1)}, \ldots, X^{(p)})$, via la modélisation de $\mathbb{E}(Y|X)$, mais :

- $\mathbb{E}(Y|X)$ s'interprète différemment selon les situations (Y binaire ou $Y \in \{A_1, \ldots, A_k\}$ ou $Y \in \mathbb{N}$), cf la suite.
- Dans tous ces cas le modèle linéaire $\mathbb{E}(Y|X) = X'\beta$ ne convient pas.
- On va modéliser $\mathbb{E}(Y|X)$ autrement : via un **modèle linéaire généralisé**.
- Comme en régression linéaire, on s'intéressa alors à l'effet spécifique d'un régresseur donné, *toutes choses égales par ailleurs.*
- Et selon les cas, on s'intéressera plutôt à l'explication ou à la prédiction.

Trois cas fondamentaux

On détaille les enjeux de la modélisation de $\mathbb{E}(Y|X)$ pour trois cas fondamentaux :

- Cas 1 : Y binaire
- Cas 2 : $Y \in \{A_1, \ldots, A_k\}$ (variable qualitative générale)
- Cas 3 : $Y \in \mathbb{N}$ (variable de comptage)

Cas 1 : Y binaire

On suppose, sans perte de généralité, que $Y \in \{0, 1\}$.

Si Y modélise l'appartenance à une catégorie, disons A, cela revient à dire que l'on s'intéresse à la variable $Y = \mathbf{1}_A$.

La loi de Y sachant $X = x$ est entièrement déterminé par

$$p(x) = \mathbb{P}(Y = 1|X = x),$$

car on en déduit $\mathbb{P}(Y = 0|X = x) = 1 - p(x)$.

En fait $Y|X = x$ est une **loi de Bernoulli** de paramètre $p(x)$ et

$$\mathbb{E}(Y|X = x) = p(x).$$

En particulier :

$$p(x) \in [0, 1].$$

On a

$$\mathbb{E}(Y|X=x) = \mathbb{P}(Y=1|X=x) = p(x) \in [0,1].$$

Comment modéliser $p(x)$?

- $p(x) = x'\beta$ (pour un certain $\beta \in \mathbb{R}^p$ à estimer) est à proscrire car c'est à valeurs dans $\mathbb{R}$ et non $[0,1]$.
- On peut par contre proposer un modèle du type

$$p(x) = f(x'\beta)$$

où f est une fonction de $\mathbb{R}$ dans $[0,1]$.

Exemples de choix possibles pour f : ... (on y reviendra).

- Cette approche permet de construire un modèle cohérent, et qui ne dépend que d'un simple paramètre β, dans l'esprit de la régression linéaire : c'est un modèle de régression linéaire *généralisé*.

Cas 2 : $Y \in \{A_1, \ldots, A_k\}$ (variable qualitative générale)

Si Y représente l'appartenance à k classes différentes $A_1, \ldots, A_k$, sa loi est déterminée par les probabilités

$$p_j(x) = \mathbb{P}(Y \in A_j | X = x), \quad \text{pour} \quad j = 1, \ldots, k$$

avec la contrainte $\sum_{j=1}^{k} p_j(x) = 1$. (Si $k = 2$, c'est le cas précédent.)

On encode Y numériquement grâce au "one-hot encoding", de telle sorte que Y devient le vecteur $Y = (\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k})$, et l'on a :

$$\mathbb{E}(Y|X = x) = \begin{pmatrix} p_1(x) \\ \vdots \\ p_k(x) \end{pmatrix}.$$

Il suffit de modéliser $p_1(x), \ldots, p_{k-1}(x)$, car $p_k(x) = 1 - \sum_{j=1}^{k-1} p_j(x)$.
Comme dans le cas binaire, on peut proposer :

$$p_j(x) = f(x'\beta_j), \quad j = 1, \ldots, k-1, \quad \text{avec } f : \mathbb{R} \to [0, 1].$$

Il y aura donc $k - 1$ paramètres inconnus à estimer, chacun dans $\mathbb{R}^p$.

Cas 3 : $Y \in \mathbb{N}$ (variable de comptage)

Si Y est à valeurs entières, on a pour tout x

$$\mathbb{E}(Y|X = x) \geq 0.$$

Ici aussi le modèle $\mathbb{E}(Y|X = x) = x'\beta$ est à proscrire car le résultat n'est pas nécessairement positif.

Un choix plus cohérent est

$$\mathbb{E}(Y|X = x) = f(x'\beta)$$

où f est une fonction de $\mathbb{R}$ dans $[0, +\infty[$.

Exemple de choix possible pour f : la fonction exponentielle.

Soit g une fonction strictement monotone, appellée **fonction de lien**.

Un modèle linéaire généralisé (GLM) établit une relation du type

$$g(\mathbb{E}(Y|X=x)) = x'\beta$$

ou, ce qui revient au même

$$\mathbb{E}(Y|X=x) = g^{-1}(x'\beta)$$

(Dans les exemples précédents, g^{-1} était noté f).

On suppose généralement en plus que

- la loi de $Y|X$ fait partie de la famille exponentielle, de paramètre β.

  $\longrightarrow$ cela permet de calculer la vraisemblance (car on connait toute la loi et non seulement son espérance) et facilite l'inférence.

- Dans un modèle GLM, le but est d'estimer $\beta \in \mathbb{R}^p$.
- Pour cela, à partir de n observations indépendantes de (Y, X), on utilisera le maximum de vraisemblance (la loi de $Y|X$ étant connue à β près).
- La fonction de lien g n'est pas à estimer : on la choisit selon la nature de Y. Pour une loi $Y|X$ donnée, il y a un choix naturel, cf la suite.
- Des outils d'inférence et de diagnostics sont disponibles (comme en régression linéaire).
- Parmi les variables explicatives $X^{(1)}, \ldots, X^{(p)}$, on supposera souvent que $X^{(1)} = 1$ pour tenir compte de la présence d'une constante. Ainsi

$$X'\beta = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} = \beta_1 + \beta_2 X^{(2)} + \cdots + \beta_p X^{(p)}.$$

(On indice parfois différemment pour écrire $\beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$)

On le retrouve en prenant la fonction de lien identité $g(t) = t$.
Alors

$$\mathbb{E}(Y|X = x) = g^{-1}(x'\beta) = x'\beta.$$

Pour le modèle linéaire Gaussien, on suppose en plus que $Y|X$ suit une loi Gaussienne de variance σ^2 (il s'agit bien d'une loi de la famille exponentielle).

Dans ce contexte,

$$Y|X \sim \mathcal{N}(X'\beta, \sigma^2).$$

La régression linéaire est donc un cas particulier des modèles GLM.

Exemple 2 : le cas binaire $Y \in \{0, 1\}$

Dans ce cas $Y|X$ est une loi de Bernoulli (qui est dans la famille exponentielle).

La fonction de lien g doit vérifier

$$\mathbb{E}(Y|X = x) = g^{-1}(x'\beta)$$

où g^{-1} est une fonction de $\mathbb{R}$ dans $[0, 1]$. On a alors

$$Y|X \sim \mathcal{B}(g^{-1}(X'\beta)).$$

Choix possible pour g^{-1} : n'importe quelle fonction de répartition d'une loi continue sur $\mathbb{R}$.

Choix standard pour g^{-1} : la fdr d'une loi logistique

$$g^{-1}(t) = \frac{e^t}{1 + e^t} \quad \Leftrightarrow \quad g(t) = \ln\left(\frac{t}{1 - t}\right) = \text{logit}(t)$$

Ce choix conduit au **modèle logistique**, le plus important de ce chapitre.

Exemple 3 : si $Y \in \mathbb{N}$

La fonction de lien $g(t) = \ln(t) \Leftrightarrow g^{-1}(t) = e^t$ donne

$$\mathbb{E}(Y|X = x) = g^{-1}(x'\beta) = e^{x'\beta}.$$

Pour la loi de $Y|X$, définie sur $\mathbb{N}$, on suppose souvent qu'il s'agit d'une loi de Poisson (qui appartient à la famille exponentielle).

Dans ce contexte,

$$Y|X \sim \mathcal{P}(e^{X'\beta}).$$

Résumé

Il y a 2 choix à faire pour mettre en place un modèle GLM :

1. La loi de $Y|X$;
2. La fonction de lien g définissant $\mathbb{E}(Y|X) = g^{-1}(X'\beta)$.

Le second choix est lié au premier.

Les 3 cas usuels :

Y binaire ($Y \in \{0, 1\}$) :

$\rightarrow$ Aucun choix pour la loi de $Y|X$: c'est une loi de Bernoulli

$\rightarrow$ Il reste à choisir g : par défaut $g = \textit{logit}$

Y prend plusieurs modalités ($Y \in \{A_1, \ldots, A_k\}$) :

$\rightarrow$ Aucun choix pour la loi de $Y|X$: c'est une loi multi-Bernoulli

$\rightarrow$ Il reste à choisir g (pour chaque modalité) : par défaut $g = \textit{logit}$

Y est à valeurs dans $\mathbb{N}$ ($Y \in \mathbb{N}$) :

$\rightarrow$ pour la loi de $Y|X$: Poisson (souvent) ou binomiale négative ou ...

$\rightarrow$ choix de g : par défaut $g = \ln$

4 Modèles linéaires généralisés

- Généralité sur les GLM (generalized linear models)
- Le modèle logistique pour Y binaire
  - La fonction logit comme fonction de lien
  - Enjeux du modèle logistique
  - Interprétation du modèle
  - Estimation des paramètres
  - Tests et intervalles de confiance
  - Déviance, tests et choix de modèles
  - Classification
- Modèles pour données catégorielles
- Modèles pour données de comptage

Le modèle logistique concerne la modélisation d'une variable Y **binaire**.
Donc dans cette partie :

- Y est une variable binaire $Y \in \{0, 1\}$
- $X = (X^{(1)}, \ldots, X^{(p)})$ sont p variables explicatives.

Pour rappel, dans ce cas :

- $Y|X = x$ est une loi de Bernoulli de paramètre $p(x) = \mathbb{P}(Y = 1|X = x)$.
- Il s'agit donc de modéliser $p(x) = \mathbb{E}(Y|X = x)$,
- à l'aide d'un modèle de la forme $p(x) = g^{-1}(x'\beta)$,
- pour une fonction g^{-1} strictement croissante à valeurs dans $[0, 1]$.

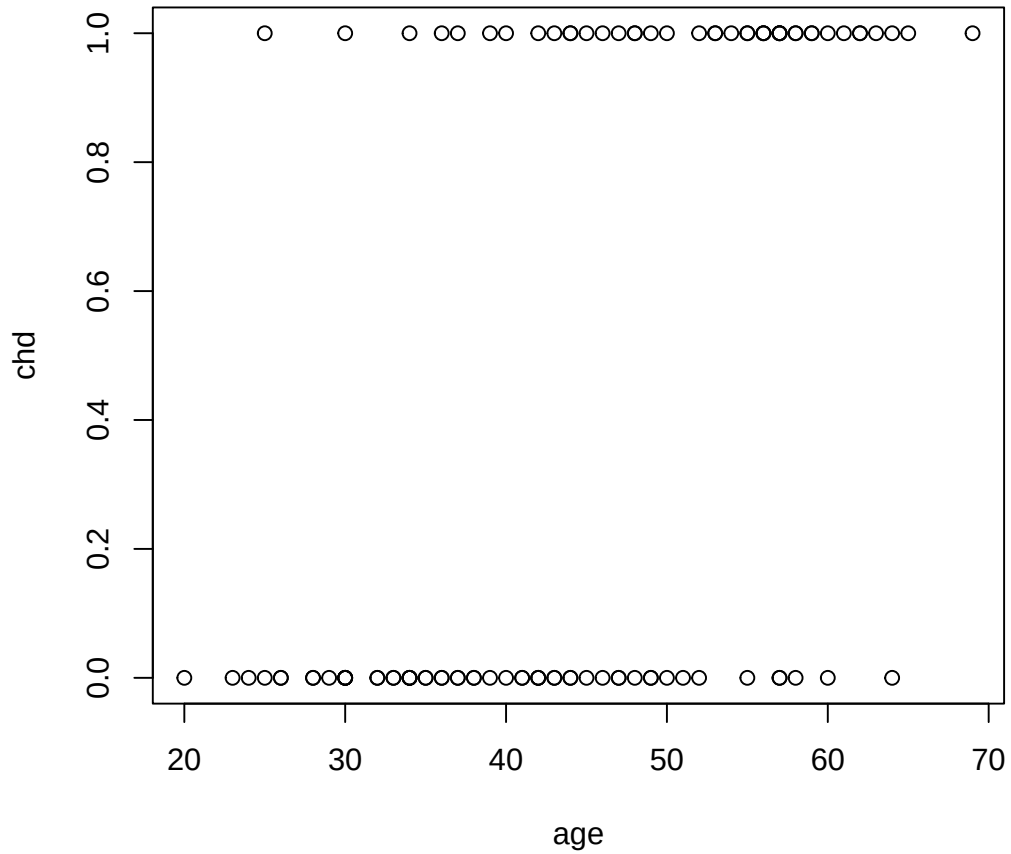
On va commencer par discuter du choix de g^{-1} (ou de g, c'est pareil).

Le modèle logistique correspond au choix naturel $g = \mathit{logit}$, cf la suite.

Exemple : présence d'une maladie coronarienne

Présence d'une maladie coronarienne ($chd = 1$ ou 0) en fonction de l'âge.
Ici : $Y = chd$ et $X = age$.

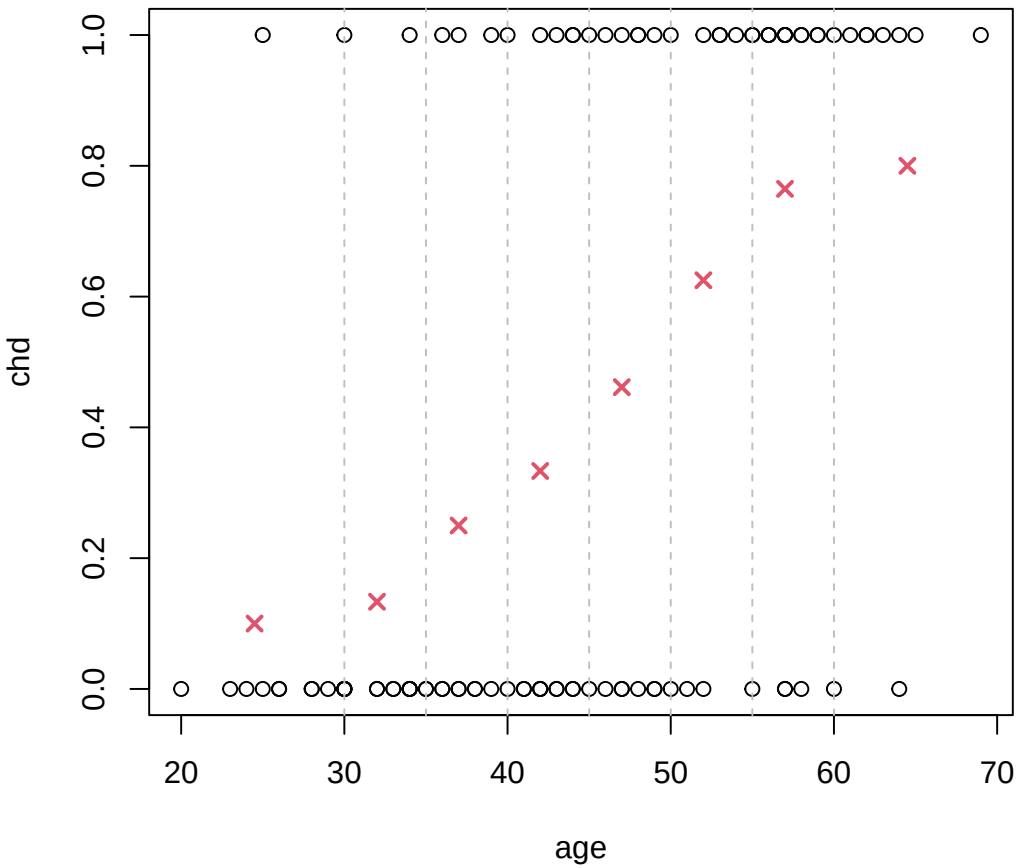
On veut estimer $p(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(chd = 1|X = x)$ pour tout x.



Exemple : estimation non-paramétrique de $p(x)$

Idée simple pour estimer $\mathbb{P}(chd = 1|X = x)$:

- on regroupe les x par classe d'âge
- on calcule la proportion de $chd = 1$ dans la classe contenant x.



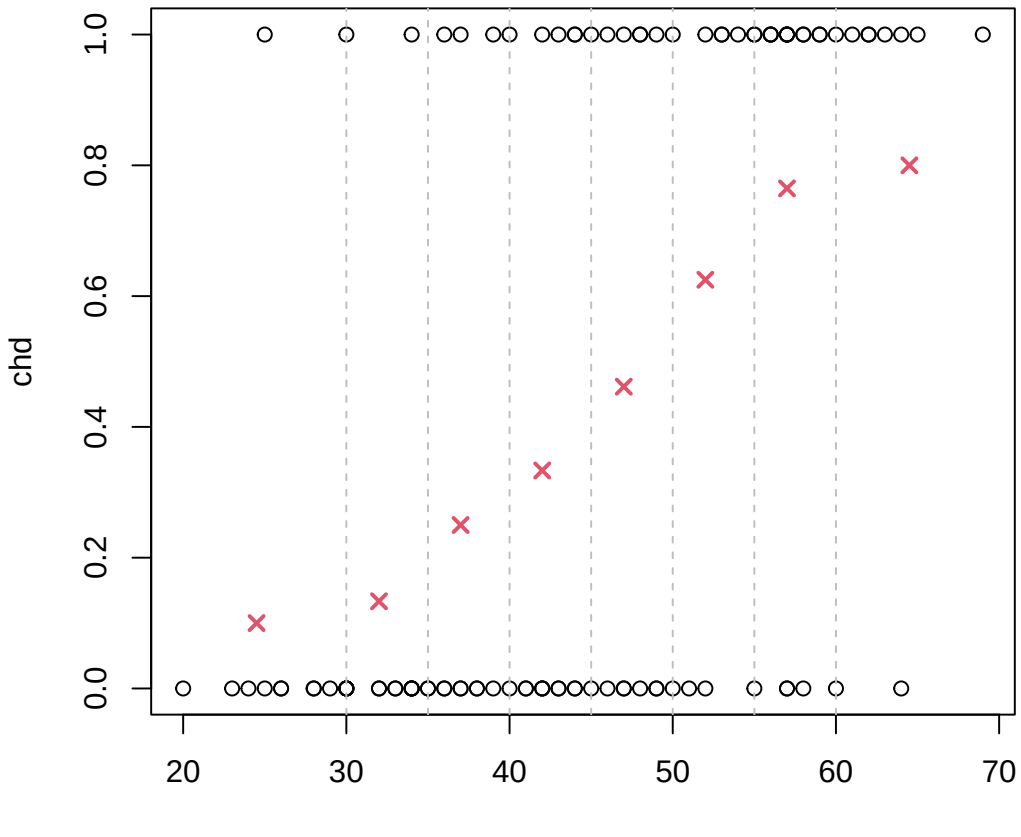
Exemple : approche GLM

On souhaite modéliser $p(x) = \mathbb{P}(chd = 1|X = x)$ par

$$p(x) = g^{-1}(\beta_0 + \beta_1 x)$$

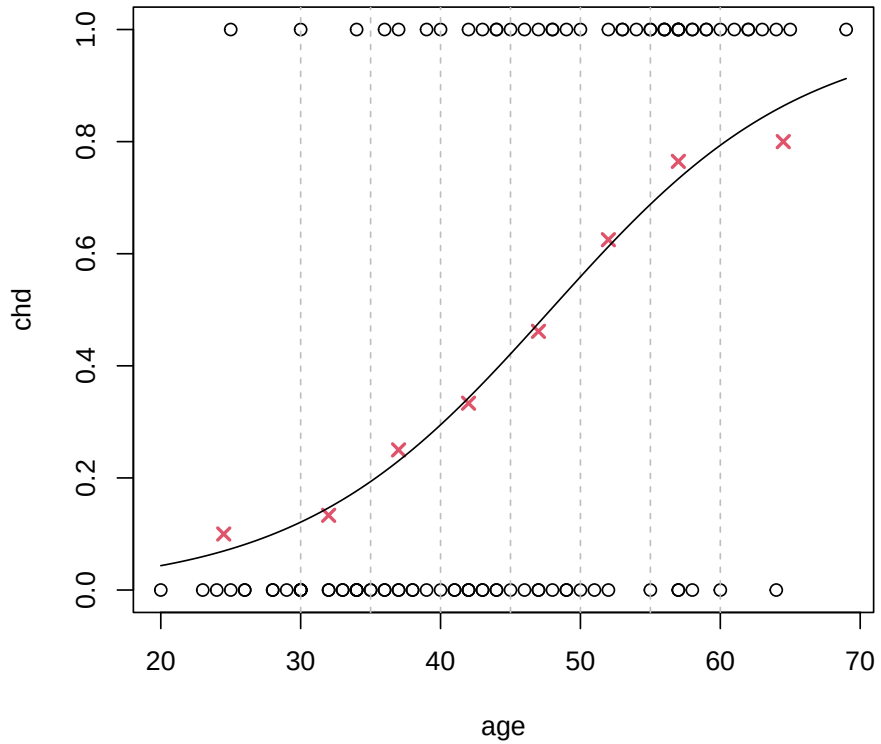
pour des paramètres bien choisis β_0 et β_1. Il faut donc que

- g^{-1} soit à valeurs dans $[0, 1]$;
- g^{-1} ait une forme de "S".



Trois choix usuels : 1) **Le modèle logit**

$$g(t) = \ln\left(\frac{t}{1-t}\right) = \text{logit}(t) \quad \Leftrightarrow \quad g^{-1}(t) = \frac{e^t}{1+e^t}.$$

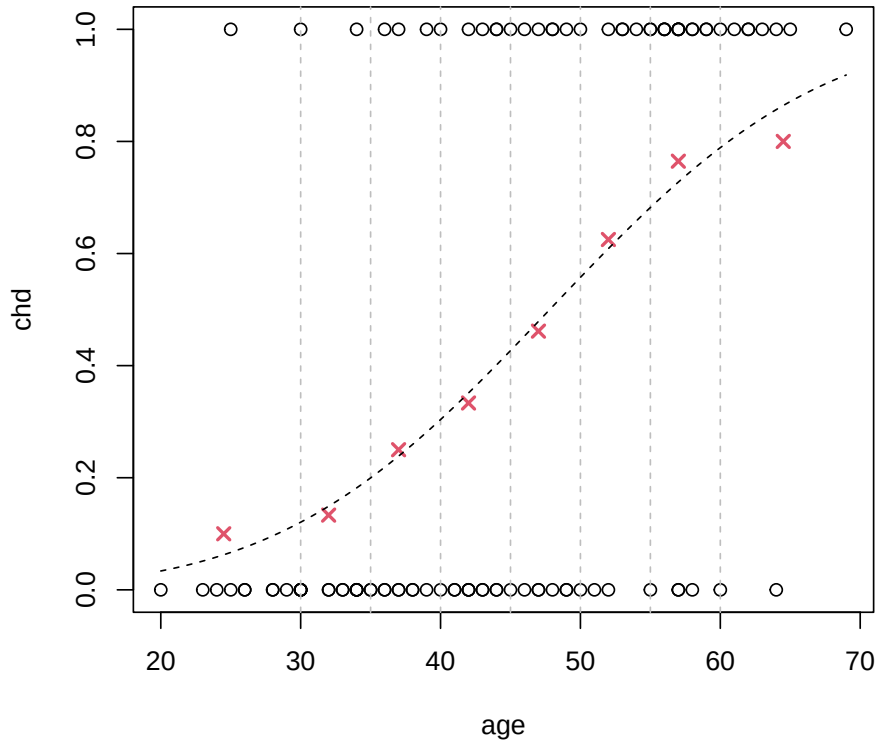


Résultat de $g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$ pour $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenus par MLE.

Exemple : fonctions de lien possibles

Trois choix usuels : 2) **Le modèle probit** où Φ : fdr d'une $\mathcal{N}(0,1)$,

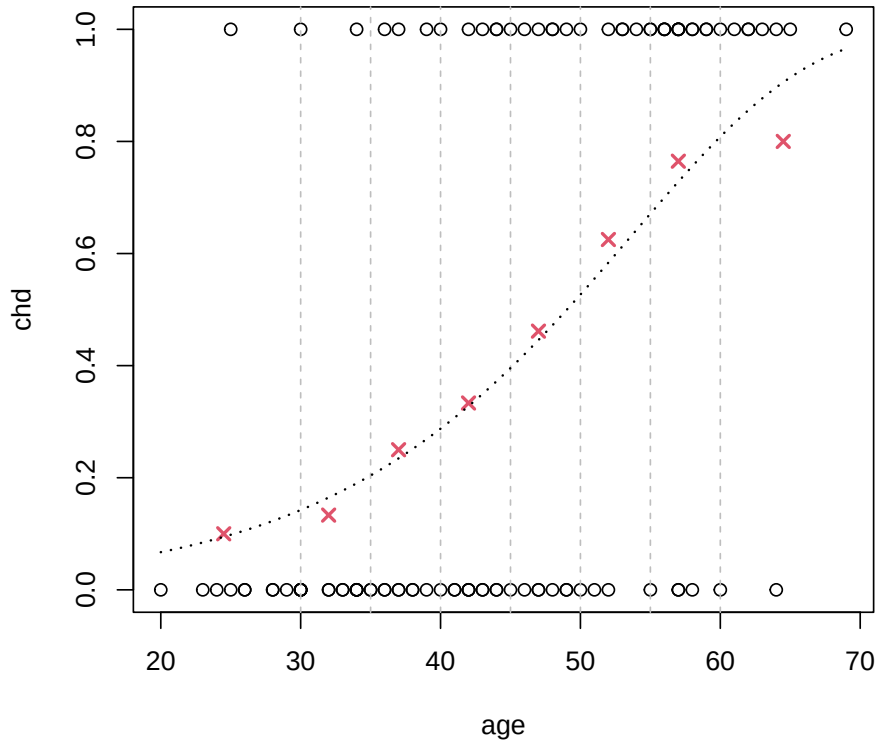
$$g(t) = \Phi^{-1}(t) \quad \Leftrightarrow \quad g^{-1}(t) = \Phi(t).$$



Résultat de $g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$ pour $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenus par MLE.

Trois choix usuels : 3) **Le modèle cloglog** (log-log complémentaire)

$$g(t) = \ln(-\ln(1-t)) \quad \Leftrightarrow \quad g^{-1}(t) = 1 - e^{-e^t}.$$

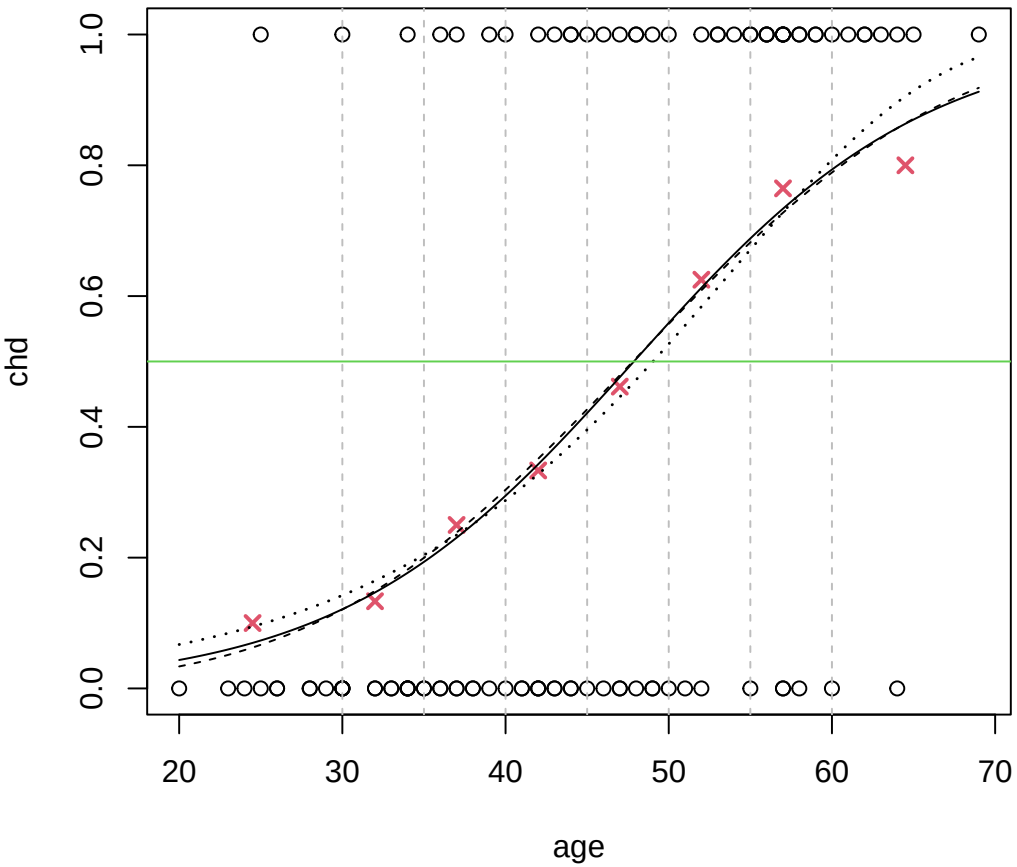


Résultat de $g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$ pour $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenus par MLE.

Exemple : fonctions de lien possibles

La différence entre les trois choix est faible :

- logit et probit donnent à peu près le même résultat
- cloglog diffère un peu et n'est pas "symétrique"



Quel fonction de lien choisir dans le cas binaire ?

Quelle fonction de lien choisir en pratique si Y est binaire ?

- Par défaut on privilégie le modèle **logit**.
- Sauf s'il y a une bonne raison de choisir autre chose (modèle probit, ou log-log complémentaire, ou log-log).
- Des détails sont fournis ci-après pour justifier ce choix par défaut.

Il est justifié lorsque la variable binaire $Y|X = x$ provient du seuillage d'une variable latente $Z(x)$ Gaussienne, i.e.,

$$(Y|X = x) = \mathbf{1}_{Z(x)\geq\tau} \quad \text{où} \quad Z(x) \sim \mathcal{N}(x'\beta, \sigma^2).$$

En effet dans ce cas

$$\mathbb{P}(Y = 1|X = x) = \mathbb{P}(Z(x) \geq \tau) = \Phi\left(\frac{x'\beta - \tau}{\sigma}\right),$$

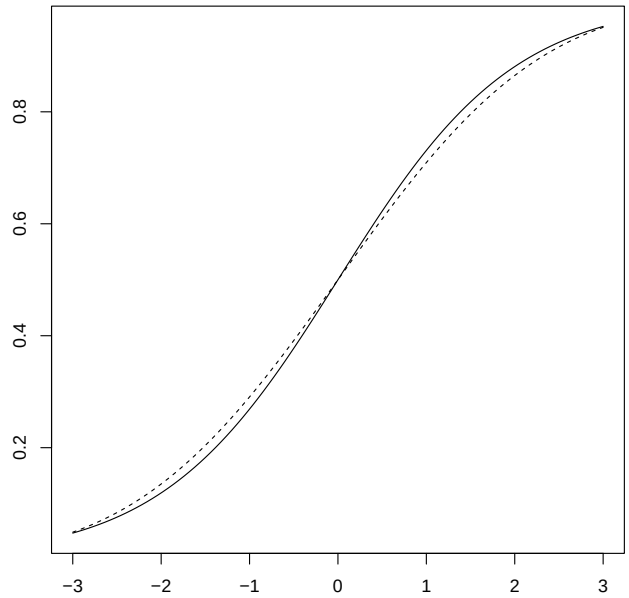
où Φ : fdr d'une $\mathcal{N}(0, 1)$, en phase avec un modèle probit.

Exemple :

- Y représente un acte d'achat, et $Z(x)$ quantifie l'utilité du bien.
- Y est un état psychologique déclaré (bonheur, dépression) et $Z(x)$ est une mesure latente, inobservée, de la satisfaction personnelle.

Détails sur le modèle probit

- Le modèle probit reste relativement prisé par les économètres...
- ...mais il tend à être remplacé par le modèle logistique.
- En effet, le modèle logit a plein d'avantages que n'a pas probit (interprétation des résultats, formules explicites,...)
- De plus, la loi logistique est très proche d'une loi Gaussienne (cf leur fdr), donc la motivation précédente pour probit reste correcte pour logit.

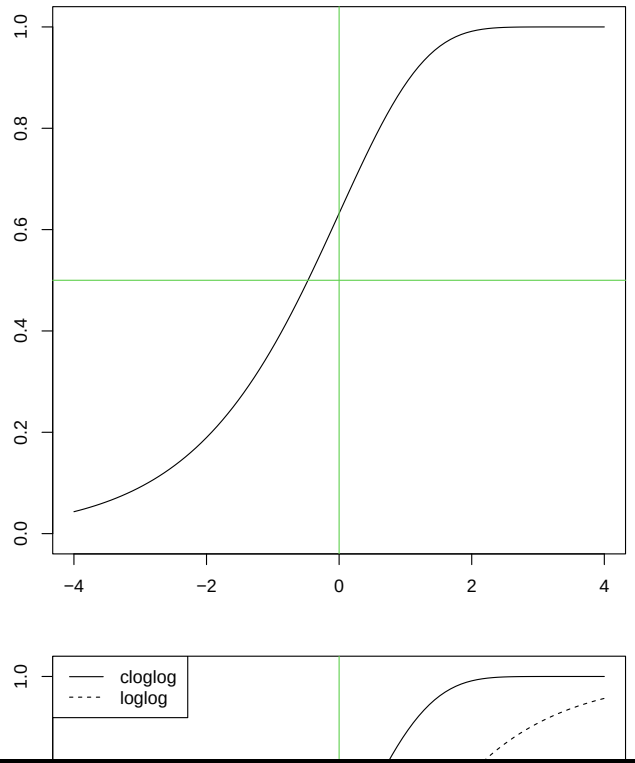


Détails sur le modèle log-log complémentaire (cloglog)

La modélisation est $p(x) = g^{-1}(x'\beta)$ avec

$$g(t) = \ln(-\ln(1-t)) \quad \Leftrightarrow \quad g^{-1}(t) = 1 - e^{-e^t}.$$

- Pas symétrique dans le sens où $g(t) \neq -g(1-t)$.
- $p(x)$ approche 0 lentement mais 1 très rapidement : modèle utile lorsque ce type de phénomène est observé.
- Si c'est l'inverse : prendre $g(t) = -\ln(-\ln(t))$ (*modèle loglog*)
- Le modèle cloglog apparait "naturellement" en analyse de survie, en lien avec le modèle de Cox (cf un autre cours)



Détails sur le modèle logit

- Il offre un outil d'interprétation très prisé : les *odds-ratios*.
- Il est plus "pratique" d'un point de vue théorique
- C'est le modèle naturel dans plein de situations.

On montrera en effet en exercice que :

- Si les deux groupes d'individus associés à $Y = 0$ et $Y = 1$ ont une loi des X qui est Gaussienne de moyenne différente, i.e. pour $m_0 \neq m_1$,

$$X|Y = 0 \sim \mathcal{N}(m_0, \Sigma) \quad et \quad X|Y = 1 \sim \mathcal{N}(m_1, \Sigma),$$

  alors $\mathbb{P}(Y = 1|X = x)$ suit un modèle logistique.
- Le résultat précédent reste vrai pour toute loi de la famille exponentielle à la place de $\mathcal{N}$.
- Le modèle logistique est celui qui maximise l'entropie, tout en s'ajustant aux données.

Conclusion pour la modélisation d'une variable binaire

Résumé :

Si Y est une variable binaire, $(Y|X = x) \sim \mathcal{B}(p(x))$.

Dans un modèle GLM pour Y, on pose $p(x) = g^{-1}(x'\beta)$ où g est :

- par défaut la fonction logit, qui est la plus naturelle ;
- éventuellement probit si on a des bonnes raisons de le justifier (mais les résultats seront similaires à logit) ;
- cloglog (ou loglog) si on a des bonnes raisons de le justifier (asymétrie forte de $p(x)$, connexion avec un modèle de Cox).

Dans la suite, nous focaliserons sur le **modèle logit**.

Définition du modèle logistique

Nous sommes donc dans la situation suivante :

- Y est une variable binaire $Y \in \{0, 1\}$
- $X = (X^{(1)}, \dots, X^{(p)})$ sont p variables explicatives.

On note $p(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$.

On a $Y|X \sim \mathcal{B}(p(X))$.

Dans le modèle logistique, on suppose que :

$$p(x) = logit^{-1}(x'\beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}},$$

où $\beta \in \mathbb{R}^p$ et $x'\beta = \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$. De façon équivalente :

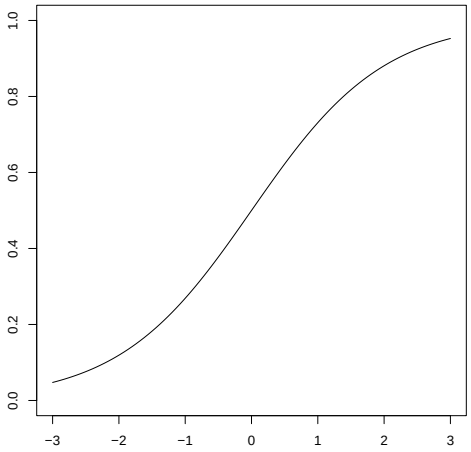
$$logit(p(x)) = \ln \frac{p(x)}{1 - p(x)} = x'\beta.$$

Enjeux

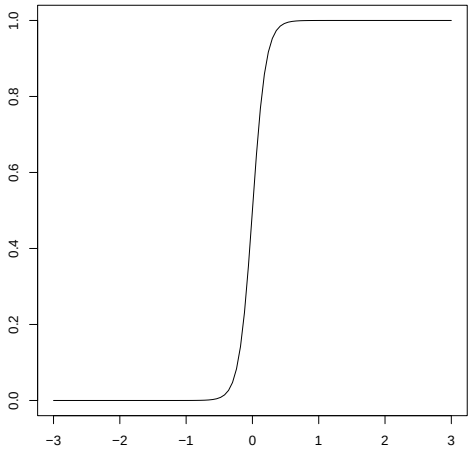
On va chercher à :

- interpréter le modèle,
- estimer β à partir d'un jeu de données,
- évaluer la qualité d'estimation,
- l'exploiter pour faire des prévisions/de la classication.

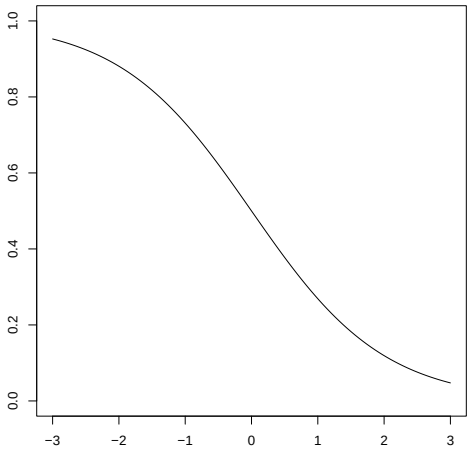
Forme de $x^{(j)} \mapsto p(x)$ en fonction de $x^{(j)}$ pour différentes valeurs de β_j



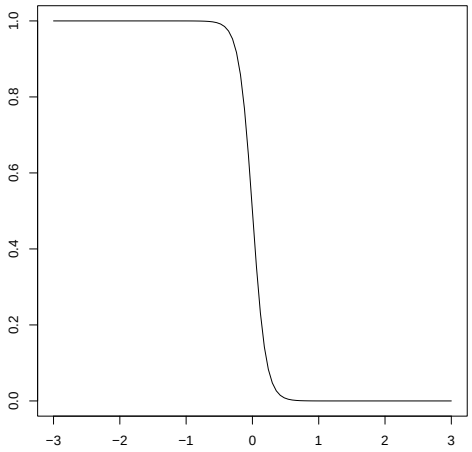
$\beta_j = 1$



$\beta_j = 10$



$\beta_j = -1$



$\beta_j = -10$

Lecture des coefficients, à retenir :

$$p(x) = logit^{-1}(x'\beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}.$$

- $x^{(j)} \to p(x)$ est croissante si $\beta_j > 0$, décroissante sinon.
- Plus $|\beta_j|$ est grand, plus le régresseur $X^{(j)}$ a un fort pouvoir de discrimination (une petite variation de $x^{(j)}$ peut occasionner une forte variation de $p(x)$).
- Attention toutefois à l'unité de mesure des régresseurs.
- Et attention à l'interprétation si le régresseur est qualitatif (comme en régression linéaire, cf l'ANOVA et l'ANCOVA).

Exemple : étude sur l'obésité aux Etats-Unis

Pour chacun des 5300 patients, on observe :

- Y : 1 si IMC>35, 0 sinon
- AGE
- DBP : pression basse (diastolique)
- SEXE : homme ou femme
- ACTIV : 1 si activité sportive intense, 0 sinon
- WALK : 1 si marche ou vélo pour aller au travail, 0 sinon
- MARITAL : statut marital (6 catégories : marié, veuf, divorcé, séparé, célibataire ou concubinage)

On cherche à modéliser $\mathbb{P}(Y = 1|X)$ où X regroupe les variables précédentes (hors Y).

Sous R, on utilise la fonction `glm` avec l'option `family=binomial`.

```
glm(Y~AGE+DBP+SEXE+ACTIV+WALK+MARITAL, family=binomial)
```

Exemple : étude sur l'obésité aux Etats-Unis

On obtient ceci :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.810240   0.294316  -9.548  < 2e-16 ***
AGE         -0.004407   0.002717  -1.622    0.105
DBP          0.017581   0.003283   5.356 8.53e-08 ***
SEXEFEMME    0.544916   0.081261   6.706 2.00e-11 ***
WALK1       -0.409344   0.095972  -4.265 2.00e-05 ***
ACTIV1      -0.789734   0.126653  -6.235 4.51e-10 ***
MARITAL2     0.070132   0.149638   0.469    0.639
MARITAL3    -0.071318   0.127510  -0.559    0.576
MARITAL4     0.188228   0.206598   0.911    0.362
MARITAL5     0.070613   0.115928   0.609    0.542
MARITAL6    -0.150165   0.157687  -0.952    0.341
```

- La lecture est semblable à celle d'un modèle de régression linéaire (mais les outils sous-jacents sont différents, cf la suite).
- On a envie de retirer la variable MARITAL du modèle.
- De plus une analyse plus poussée nous invite à intégrer un terme quadratique pour la variable AGE

Exemple : étude sur l'obésité aux Etats-Unis

On lance donc

```
glm(Y~AGE+I(AGE^2)+DBP+SEXE+WALK+ACTIV, family=binomial)
```

pour obtenir

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9564361  0.3155529 -12.538  < 2e-16 ***
AGE          0.0640837  0.0123960   5.170 2.34e-07 ***
I(AGE^2)    -0.0006758  0.0001260  -5.364 8.14e-08 ***
DBP          0.0121546  0.0033775   3.599  0.00032 ***
SEXEFEMME    0.5155651  0.0776229   6.642 3.10e-11 ***
WALK1       -0.4042257  0.0913195  -4.426 9.58e-06 ***
ACTIV1      -0.6573558  0.1150226  -5.715 1.10e-08 ***
```

Par exemple, selon ce modèle, pour un homme qui n'a pas d'activité sportive intense et ne va pas au travail à pied ou à vélo :

$$\mathbb{P}(Y = 1|AGE, DBP) = logit^{-1}(-3.95 + 0.064\, AGE - 0.00068\, AGE^2 + 0.0122\, DBP).$$

Pour le même profil mais avec ACTIV="1" :

$$\mathbb{P}(Y = 1|AGE, DBP) = logit^{-1}(-3.95 + 0.064\, AGE - 0.00068\, AGE^2 + 0.0122\, DBP - 0.657).$$

Si p est la probabilité d'un événement A, alors sa cote (odds) vaut :

$$odds = \frac{p}{1-p}.$$

C'est la cote au sens des paris : par exemple 3 contre 1 signifie que pour 3 personnes pariant sur A, 1 personne parie sur B. Donc un parieur pris au hasard a une probabilité 3/4 de parier sur A ($p = 3/4$ et $odds = 3$).

De même l'odds (ou cote) d'obtenir $Y = 1$ sachant $X = x$ est

$$odds(x) = \frac{p(x)}{1-p(x)}$$

où $p(x) = \mathbb{P}(Y = 1|X = x)$.

Si deux individus présentent les caractéristiques x_1 et x_2 respectivement, on appelle odds-ratio entre x_1 et x_2

$$OR(x_1, x_2) = \frac{odds(x_1)}{odds(x_2)} = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}}$$

NE PAS CONFONDRE ODDS RATIO ET RAPPORT DE PROBABILITES

Seule exception possible : si $p(x_1)$ et $p(x_2)$ sont très petits car alors $1 - p(x_1) \approx 1$ et $1 - p(x_2) \approx 1$, de sorte que $OR(x_1, x_2) \approx p(x_1)/p(x_2)$

NE PAS CONFONDRE ODDS RATIO ET RAPPORT DE PROBABILITES

Toutefois il reste que (cf tableau) :

$$OR(x_1, x_2) > 1 \iff \frac{p(x_1)}{p(x_2)} > 1$$
$$OR(x_1, x_2) < 1 \iff \frac{p(x_1)}{p(x_2)} < 1$$
$$OR(x_1, x_2) = 1 \iff \frac{p(x_1)}{p(x_2)} = 1$$

Attention néanmoins à la significativité (tests, IC, cf. la suite).

$OR(x_1, x_2)$ accentue les différences par rapport à $p(x_1)/p(x_2)$:

$$OR(x_1, x_2) > 1 \iff OR(x_1, x_2) > \frac{p(x_1)}{p(x_2)}$$
$$OR(x_1, x_2) < 1 \iff OR(x_1, x_2) < \frac{p(x_1)}{p(x_2)}$$

Utilisation des odds-ratios

Une régression logistique sert le plus souvent à comparer le comportement de deux individus vis-à-vis de la variable d'intérêt.

Exemples :

- probabilité d'achat selon qu'on a ou non fait l'objet d'une promotion personnalisée, *toutes choses égales par ailleurs* ;
- pour un véhicule donnée, probabilité d'avoir une panne selon l'ancienneté, *toutes choses égales par ailleurs* ;
- probabilité de guérison selon le traitement utilisé, *toutes choses égales par ailleurs* ;
- ...

Dans le cadre d'une régression logistique,

$$odds(x) = \frac{p(x)}{1-p(x)} = \exp(x'\beta).$$

Ainsi :

$$OR(x_1, x_2) = \frac{odds(x_1)}{odds(x_2)} = \exp((x_1 - x_2)'\beta)$$

donc si les deux individus ne diffèrent que par le régresseur j

$$OR(x_1, x_2) = \exp(\beta_j(x_1^{(j)} - x_2^{(j)})).$$

et si le régresseur j est binaire ($x_1^{(j)} = 1$ alors que $x_2^{(j)} = 0$) :

$$OR(x_1, x_2) = \exp(\beta_j).$$

La phrase qui résume le tout :

Dans un modèle de régression logistique, β_j s'interprète comme le logarithme de l'odds-ratio entre deux individus différents d'une quantité 1 sur le régresseur j, **toutes choses égales par ailleurs**.

$$\boxed{\text{En bref : } \exp(\beta_j) = OR(x^{(j)} + 1, x^{(j)})}$$

Si le régresseur j est binaire (absence ou présence d'une certaine caractéristique) : $\exp(\beta_j)$ est simplement l'OR entre la présence ou non de cette caractéristique, **toutes choses égales par ailleurs**.

Point de vue professionnel :

- En pratique, on est souvent intéressé en premier lieu par l'impact d'une caractéristique particulière sur $p(x)$.
- De ce point de vue, le modèle logistique rend l'utilisation des odds-ratio particulièrement agréable (et explique en partie la popularité du modèle logistique).

Exemple : étude sur l'obésité aux Etats-Unis

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9564361  0.3155529 -12.538  < 2e-16 ***
AGE          0.0640837  0.0123960   5.170 2.34e-07 ***
I(AGE^2)    -0.0006758  0.0001260  -5.364 8.14e-08 ***
DBP          0.0121546  0.0033775   3.599  0.00032 ***
SEXEFEMME    0.5155651  0.0776229   6.642 3.10e-11 ***
WALK1       -0.4042257  0.0913195  -4.426 9.58e-06 ***
ACTIV1      -0.6573558  0.1150226  -5.715 1.10e-08 ***
```

- L'Odds Ratio correspondant à la pratique ou non d'une activité sportive intense vaut, toutes choses égales par ailleurs :

$$\exp(-0.657) \approx 0.52$$

La cote de la survenue de l'obésité diminue donc de moitié pour les individus pratiquant une activité sportive intense.
(La cote et non la probabilité !)

- L'OR pour un écart de pression diastolique de $+20$ vaut :

$$\exp(0.0121546 \times 20) \approx 1.28$$

La cote de la survenue de l'obésité augmente donc de 28% dans ce cas.

On observe n réalisations i.i.d. (Y_i, X_i) où $Y_i \in \{0, 1\}$ et $X_i \in \mathbb{R}^p$.

On note $p(x_i) = \mathbb{P}(Y_i = 1 | X_i = x_i)$.

On suppose le modèle logistique : pour tout i,

$$p(x_i) = logit^{-1}(x_i'\beta) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}},$$

où $\beta = (\beta_1, \ldots, \beta_p)$ et $x_i'\beta = \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)}$.

On va estimer β en maximisant la vraisemblance.

On notera $p_\beta(x_i)$ pour souligner la dépendance de $p(x_i)$ en β.

Calcul de la vraisemblance

Pour tout i, $Y_i|(X_i = x_i)$ suit la loi $\mathcal{B}(p_\beta(x_i))$. Donc

$$\mathbb{P}(Y_i = y_i|X_i = x_i) = p_\beta(x_i)^{y_i}(1 - p_\beta(x_i))^{1-y_i}, \quad \text{pour tout } y_i \in \{0, 1\}.$$

Par indépendance, on obtient donc la vraisemblance de l'échantillon

$$V(\beta, y_1, \ldots, y_n|x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\beta(x_i)^{y_i}(1 - p_\beta(x_i))^{1-y_i}.$$

En passant au log et en remplaçant $p(x_i)$ par son expression, on obtient la log-vraisemblance :

$$L(\beta, y_1, \ldots, y_n|x_1, \ldots, x_n) = \ln V() = \sum_{i=1}^{n} \left(y_i\, x_i'\beta - \ln(1 + e^{x_i'\beta}) \right).$$

L'EMV $\hat{\beta}$, s'il existe, annule le gradient de L par rapport à β. Ce gradient vaut

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} x_i \left(y_i - \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right).$$

Il s'agit donc de résoudre un système de p équations à p inconnues.

Mais la solution n'est pas explicite...

Il faut recourir à des méthodes numériques pour résoudre le système (algorithme de Newton-Raphson).

Remarque : Il s'agit d'une situation classique dès lors qu'on utilise des modèles statistiques avancés : on a souvent recours à des algorithmes d'optimisation.

Question pratique importante : la solution existe-t-elle ? Est-elle unique ?

Soit $\mathbf{X}$ la matrice de design (dont les lignes sont les vecteurs x_i').

Proposition (Unicité)

Si $rg(\mathbf{X}) = p$, alors l'EMV dans un modèle logistique, s'il existe, est unique.

Preuve : Il suffit de montrer que L est strictement concave en β.
Pour cela on calcule la matrice Hessienne de L (détails au tableau) :

$$\frac{\partial^2 L}{\partial \beta^2} = -\sum_{i=1}^{n} p_\beta(x_i)(1 - p_\beta(x_i))\ x_i x_i'.$$

Elle est semi-définie négative. De plus, pour tout $u \in \mathbb{R}^p$,

$$u' \frac{\partial^2 L}{\partial \beta^2} u = 0 \Leftrightarrow u' x_i x_i' u = 0 \text{ pour tout } i, \text{ car } p_\beta(x_i)(1 - p_\beta(x_i)) > 0$$

$$\Leftrightarrow x_i' u = 0 \text{ pour tout } i$$

$$\Leftrightarrow \mathbf{X}u = 0$$

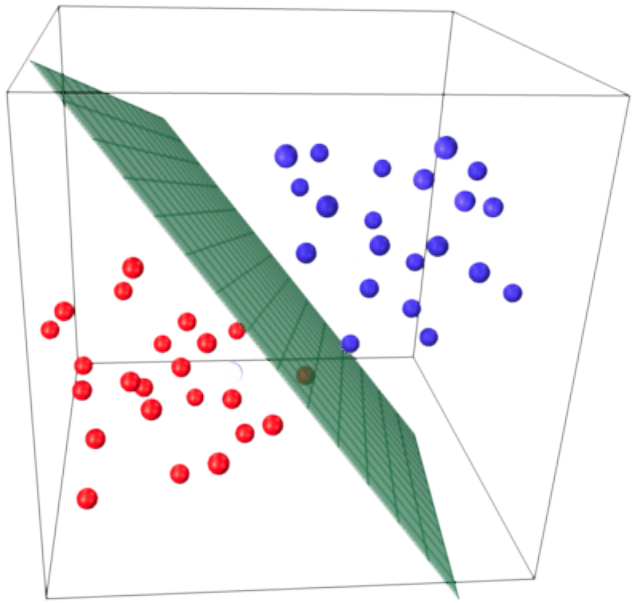
$$\Leftrightarrow u = 0 \quad \text{car } rg(\mathbf{X}) = p.$$

Ainsi, pour tout $u \neq 0$, $u' \frac{\partial^2 L}{\partial \beta^2} u < 0$: la matrice Hessienne est définie négative et donc L est strictement concave.

A propos de l'existence de l'EMV

Bien que L soit strictement concave, son maximum peut survenir à l'infini (penser à la fonction ln), auquel cas $\hat{\beta}$ n'existe pas.

Cela arrive s'il y a **non-recouvrement**, cad séparation par un hyperplan des x_i pour lesquels $y_i = 0$ (les bleus) et de ceux pour lesquels $y_i = 1$ (les rouges).



Mathématiquement, il y a non-recouvrement s'il existe $\alpha \in \mathbb{R}^p$ tel que

$$\begin{cases} \text{pour tout } i \text{ tel que } y_i = 0, & \alpha' x_i \geq 0, \\ \text{pour tout } i \text{ tel que } y_i = 1, & \alpha' x_i \leq 0. \end{cases}$$

Proposition (admise)

En cas de non-recouvrement, l'estimateur $\hat{\beta}$ n'existe pas, dans le sens où $L(\beta)$ est maximal lorsque $\|\beta\| \to \infty$ (dans une ou plusieurs directions).

Conséquences du non-recouvrement :

- Pour tout x, $\hat{p}(x) = 0$ ou 1, selon la position de x par rapport à l'hyperplan séparateur.
- Néanmoins, il y a une "zone morte" au milieu des 2 nuages de points, car l'hyperplan séparateur n'est pas forcément unique.
- Au delà de cette zone morte, la classification est très simple.
- Mais aucune interprétation du modèle n'est possible (les OR valent 0 ou $+\infty$).

On dit qu'il y a recouvrement lorsqu'aucun hyperplan ne peut séparer les points rouges des points bleus.

Proposition (admise)

Si $rg(\mathbf{X}) = p$ et qu'il y a recouvrement, alors l'EMV existe et est unique.

- Sous ces conditions, on peut donc rechercher l'EMV par l'algorithme de Newton-Raphson.
- La convergence de l'algorithme vers l'EMV est alors garanti puisque
  1) le maximum existe,
  2) la fonction à optimiser est strictement concave et il n'y a donc aucun maximum local, uniquement un maximum global.

Soit $\mathbf{X}$ la matrice de design (dont les lignes sont les vecteurs x_i').
Soit $J_n(\beta)$ la matrice d'information de Fisher : $J_n(\beta) = -\mathbb{E}\left(\frac{\partial^2 L}{\partial \beta^2}(\beta)\middle|\mathbf{X}\right)$

Proposition (admise)

Dans le modèle de régression logistique, si

i) *la loi des régresseurs* $(X_1, \ldots, X_p)$ *est à support compact,*

ii) $rg(\mathbf{X}) = p$,

iii) *la plus petite valeur propre* $\mathbf{X}'\mathbf{X}$ *tend vers l'infini avec n,*

alors

1. *l'estimateur du maximum de vraisemblance* $\widehat{\beta}$ *est consistant ;*
2. $$J_n(\beta)^{1/2}\left(\widehat{\beta} - \beta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p)$$

   où I_p *la matrice identité de taille p.*

- Sous ces conditions, l'EMV existe donc pour n suffisamment grand. En fait, si les données suivent un modèle logistique, il y a forcément recouvrement lorsque n est grand (cf tableau).
- Il est asymptotiquement efficace (=variance asymptotique minimale)
- La matrice d'information de Fisher $J_n(\beta)$ se calcule, cf la suite.
- On va pouvoir s'appuyer sur la normalité asymptotique pour faire des tests et construire des ICs, cf la suite.
- Les hypothèses sont raisonnables en pratique : ii) et iii) se retrouvent en régression linéaire, et i) est une manière simple de s'assurer que la loi des régresseurs est concentrée (par ailleurs cette hypothèse peut-être allégée pour intégrer des lois à support non compact, mais concentrée, comme la loi normale).
- La preuve revient aux propriétés asymptotiques du maximum de vraisemblance dans une famille exponentielle (ici la loi de Bernoulli), dans l'esprit du cours de Statistique Inférentielle.

  Pour plus de détails : cf chapitre V.5. de Antoniadis, A. Berruyer J. et Carmona R. Régression non linéaire et applications.

Différence avec la régression linéaire

En régression linéaire, pour l'estimation de β :

- La formule de $\hat{\beta}$ est explicite : $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$;
- Son espérance et sa variance sont explicites ;
- Dans le modèle Gaussien ($Y|X$ Gaussien), la loi de $\hat{\beta}$ est explicite,
- ce qui permet de construire des tests exacts (Student, Fisher).
- Si le modèle n'est pas Gaussien, ces tests sont valables asymptotiquement.

Dans le modèle logistique, pour l'estimation de β :

- Pas de formule explicite pour $\hat{\beta}$, la solution est obtenue numériquement ;
- On ne connait ni le biais, ni la variance de $\hat{\beta}$;
- La loi de $Y|X$ est simple (une Benoulli), mais on ne connait pas la loi de $\hat{\beta}$.
- **On connait uniquement sa loi asymptotique**.
- Tous les tests que nous allons construire seront donc asymptotiques.

Rappel : sous "de bonnes conditions",

$$J_n(\beta)^{1/2}\left(\widehat{\beta}-\beta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p)$$

où $J_n(\beta)$ est la matrice d'information de Fisher.

À partir de ce résultat, on va pouvoir calculer des régions asymptotiques de confiance pour β et construire des tests asymptotiques.

Il faut pour cela connaitre $J_n(\beta)$ et pouvoir l'estimer.

Par définition

$$J_n(\beta) = -\mathbb{E}\left(\frac{\partial^2 L}{\partial \beta^2}(\beta)\middle|\mathbf{X}\right)$$

où L est la log-vraisemblance du modèle.

Cette définition est valable car nous sommes dans un modèle régulier (la loi conditionnelle des Y_i est une loi de Bernoulli).

Le calcul de la Hessienne de L a déjà été fait (pour montrer l'unicité de l'EMV). On obtient

$$J_n(\beta) = \sum_{i=1}^{n} p_\beta(x_i)(1 - p_\beta(x_i))\ x_i x_i'.$$

On peut écrire de façon équivalente

$$J_n(\beta) = \mathbf{X}' W_\beta \mathbf{X}$$

où $\mathbf{X}$ est la matrice de design et W_β est la matrice diagonale

$$W_\beta = \begin{pmatrix} p_\beta(x_1)(1-p_\beta(x_1)) & & 0 \\ & \ddots & \\ 0 & & p_\beta(x_n)(1-p_\beta(x_n)) \end{pmatrix}.$$

- Pour estimer $J_n(\beta)$, on remplace simplement β par l'EMV $\hat{\beta}$
- Sous les mêmes hypothèses de régularité, on peut montrer que

$$J_n(\widehat{\beta})^{1/2}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p).$$

(c'est dans l'esprit du lemme de Slutsky, mais différent)

On se base sur la convergence

$$J_n(\widehat{\beta})^{1/2}(\widehat{\beta}-\beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p).$$

En notant $\widehat{\sigma_j}^2$ le j-ème élément diagonal de $J_n(\widehat{\beta})^{-1}$, on obtient (admis) :

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

On en déduit un intervalle de confiance pour β_j, au niveau de confiance asymptotique $1-\alpha$:

$$IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - q(1-\alpha/2)\widehat{\sigma}_j \,;\, \widehat{\beta}_j + q(1-\alpha/2)\widehat{\sigma}_j\right]$$

où $q(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

On vérifie qu'on a bien $\mathbb{P}(\beta_j \in IC_{1-\alpha}(\beta_j)) \to 1-\alpha$.

Test de significativité pour un coefficient

On souhaite tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Sous H_0 on sait que

$$\frac{\widehat{\beta}_j}{\widehat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

On en déduit une région critique au niveau asymptotique α :

$$RC_\alpha = \left\{ \frac{|\hat{\beta}_j|}{\widehat{\sigma}_j} > q(1-\alpha/2) \right\}.$$

En effet $\mathbb{P}_{H_0}(RC_\alpha) \to \alpha$.

Ce test est appelé **test de Wald**.

(Tout test statistique qui s'appuie sur la normalité asymptotique est appelé test de Wald)

En notant Φ la fdr de la loi $\mathcal{N}(0,1)$, la p-value du test vaut

$$p - value = 2\left(1 - \Phi\left(\frac{|\hat{\beta}_j|}{\widehat{\sigma}_j}\right)\right).$$

Exemple de l'obésité aux Etats-Unis

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9564361  0.3155529 -12.538  < 2e-16 ***
AGE          0.0640837  0.0123960   5.170 2.34e-07 ***
I(AGE^2)    -0.0006758  0.0001260  -5.364 8.14e-08 ***
DBP          0.0121546  0.0033775   3.599  0.00032 ***
SEXEFEMME    0.5155651  0.0776229   6.642 3.10e-11 ***
WALK1       -0.4042257  0.0913195  -4.426 9.58e-06 ***
ACTIV1      -0.6573558  0.1150226  -5.715 1.10e-08 ***
```

Dans cette sortie, pour chaque variable :

- La première colonne correspond à l'estimation $\hat{\beta}_j$
- La seconde colonne correspond à $\widehat{\sigma_j}$ (écart-type estimé de $\hat{\beta}_j$)
- La troisième colonne est la statistique de Wald $\hat{\beta}_j / \widehat{\sigma_j}$
- La quatrième colonne est la p-value associée au test de Wald
- Les étoiles témoignent de la significativité du test

Intervalle de confiance asymptotique pour un odds-ratio

On considère deux individus 1 et 2 qui diffèrent uniquement par leur régresseur j. On rappelle que dans le modèle logistique, on a alors

$$OR(x_1, x_2) = e^{\beta_j (x_1^{(j)} - x_2^{(j)})}.$$

On cherche à comparer cette valeur à 1.

L'estimation de $OR(x_1, x_2)$ est simplement

$$\widehat{OR}(x_1, x_2) = e^{\hat{\beta}_j (x_1^{(j)} - x_2^{(j)})}.$$

Exemple important : si le régresseur j est binaire avec $x_1^{(j)} = 1$ et $x_2^{(j)} = 0$, alors

$$\widehat{OR}(x_1, x_2) = e^{\hat{\beta}_j}.$$

Intervalle de confiance asymptotique pour un odds-ratio

On a vu qu'un IC asymptotique au niveau de confiance $1-\alpha$ pour $\hat{\beta}_j$ est

$$IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - q(1-\alpha/2)\widehat{\sigma}_j \, ; \, \widehat{\beta}_j + q(1-\alpha/2)\widehat{\sigma}_j\right].$$

En notant $IC_{1-\alpha}(\beta_j) = [l, r]$, cela signifie que $\mathbb{P}(l \leq \beta_j \leq r) \to 1-\alpha$.

Supposons que $x_1^{(j)} > x_2^{(j)}$. Par monotonie de la fonction exp,

$$\begin{aligned}\mathbb{P}(l \leq \beta_j \leq r) &= \mathbb{P}\left(e^{l(x_1^{(j)}-x_2^{(j)})} \leq e^{\beta_j(x_1^{(j)}-x_2^{(j)})} \leq e^{r(x_1^{(j)}-x_2^{(j)})}\right) \\ &= \mathbb{P}\left(e^{l(x_1^{(j)}-x_2^{(j)})} \leq OR(x_1,x_2) \leq e^{r(x_1^{(j)}-x_2^{(j)})}\right),\end{aligned}$$

ce qui montre qu'un IC asymptotique pour $OR(x_1, x_2)$ au niveau $1-\alpha$ est

$$IC_{1-\alpha}(OR(x_1,x_2)) = \left[e^{l(x_1^{(j)}-x_2^{(j)})}, e^{r(x_1^{(j)}-x_2^{(j)})}\right].$$

(Dans le cas où $x_1^{(j)} < x_2^{(j)}$, il suffit d'inverser les bornes.)

Test de significativité d'un odds-ratio

On souhaite généralement comparer $OR(x_1, x_2)$ à 1. Or on remarque que

$$OR(x_1, x_2) = 1 \Leftrightarrow e^{\beta_j (x_1^{(j)} - x_2^{(j)})} = 1 \Leftrightarrow \beta_j = 0.$$

- Tester $H_0 : OR(x_1, x_2) = 1$ contre $H_1 : OR(x_1, x_2) \neq 1$ revient à tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, dont une région critique au niveau asymptotique α est

$$RC_\alpha = \left\{ \frac{|\hat{\beta}_j|}{\widehat{\sigma}_j} > q(1 - \alpha/2) \right\}.$$

Néanmoins, pour les OR, on privilégie souvent des tests unilatéraux. Par exemple :

- Tester $H_0 : OR(x_1, x_2) \leq 1$ contre $H_1 : OR(x_1, x_2) > 1$ revient à tester $H_0 : \beta_j \leq 0$ contre $H_1 : \beta_j > 0$ (si $x_1^{(j)} > x_2^{(j)}$), dont une région critique au niveau asymptotique α est

$$RC_\alpha = \left\{ \frac{\hat{\beta}_j}{\widehat{\sigma}_j} > q(1 - \alpha) \right\}.$$

Supposons qu'on ait n observations au total dont

- $n(x)$ individus i présentent le jeu de caractères $x_i = x$;
- $n_1(x)$ individus i présentent le jeu de caractères $x_i = x$ et $y_i = 1$.

Le **modèle saturé** est celui qui estime $p(x)$, pour un x observé, par

$$\hat{p}_{sat}(x) = \frac{n_1(x)}{n(x)}.$$

Cas courant : Si toutes les observations sont distinctes, cad que chaque x observé ne l'est que pour un seul individu, alors pour un x observé :

$$n(x) = 1,\ n_1(x) = 0 \text{ ou } 1, \text{ et } \hat{p}_{sat}(x) = 0 \text{ ou } 1.$$

Modèle saturé

- Le modèle saturé est le modèle le plus simple à imaginer.
- Il colle parfaitement aux données.
- En revanche il ne possède aucun pouvoir explicatif (effet des régresseurs sur Y ?).
- Et il ne dit rien sur $p(x)$ si x n'est pas observé.
- Son pouvoir de généralisation est donc nul.
- Pour cette raison, le modèle saturé n'est pas vraiment un modèle...
- ...mais il va nous servir de référence en termes d'ajustement.

Autre façon de voir le modèle saturé :

- Il s'agit du modèle obtenu par maximum de vraisemblance, si on associe un paramètre par observation distincte, les paramètres étant $p_i = p(x_i)$ pour les x_i distincts (il y en a au plus n).
- Tout autre modèle peut donc être vu comme un sous-modèle du modèle saturé.

Vraisemblance du modèle saturé

Rappel : la log-vraisemblance de l'échantillon vaut, pour $y_i \in \{0, 1\}$,

$$L(y_1, \ldots, y_n | x_1, \ldots, x_n) = \sum_{i=1}^{n} y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i)).$$

On note L_{sat} la log-vraisemblance calculée pour les paramètres du modèle saturé :

- Si toutes les observations x_i sont distinctes, on a $\hat{p}_{sat}(x_i) = y_i$ avec $y_i \in \{0, 1\}$. On a ainsi
$$L_{sat} = 0.$$
Le modèle saturé est donc celui qui a la log-vraisemblance la plus grande possible : il colle parfaitement aux données (trop).
- Si les observations x_i ne sont pas distinctes, on obtient
$$L_{sat} = \sum_{x} n_1(x) \ln\left(\frac{n_1(x)}{n(x)}\right) + (n(x) - n_1(x)) \ln\left(1 - \frac{n_1(x)}{n(x)}\right)$$
où la somme parcourt l'ensemble des valeurs x prises par les x_i.

La déviance comme substitut à la SCR

La **déviance** d'un modèle mesure à quel point ce modèle *dévie* du modèle saturé (le modèle idéal au sens de la vraisemblance).

Mathématiquement, la déviance d'un modèle (logistique ou autre) vaut :

$$D = 2\left(L_{sat} - L_{mod}\right)$$

où L_{mod} désigne la log-vraisemblance pour les paramètres du modèle.
On a toujours $D \geq 0$.

Cas courant : Si toutes les observations sont distinctes, $L_{sat} = 0$ donc

$$D = -2L_{mod}.$$

La déviance joue le rôle de la SCR d'un modèle linéaire : plus la déviance est élevée, moins le modèle est bien ajusté aux données.

Sous R : La déviance retournée est $-2L_{mod}$: le terme L_{sat} est donc omis.

Comme en régression linéaire, on aimerait tester

$$H_0 : R\beta = 0 \quad \text{contre} \quad H_1 : R\beta \neq 0$$

où R est une matrice de contraintes de taille (q, p).

Pour rappel, selon le choix de R cela permet :

- de tester le minimum : y-a-t-il au moins un régresseur pertinent ?
- de comparer des modèles emboités ;
- de s'intéresser à la significativité collective d'une famille de régresseurs, par exemple :
  - les régresseurs rendant compte des conditions familiales dans une étude épidémiologique contenant également d'autres types de variables ;
  - un régresseur qualitatif dont les modalités apparaissent séparément dans l'écriture du modèle.

Test de contraintes linéaires : procédure de test

En GLM, plusieurs procédures de tests répondent au problème.

1. Le test de Wald, basé sur la normalité asymptotique de $\hat{\beta}$, qui généralise celui vu pour tester $\beta_j = 0$ contre $\beta_j \neq 0$.
2. Le test du rapport de vraisemblance, appelé dans ce contexte test de déviance.
3. Le test du score, basé sur le comportement du gradient de la log-vraisemblance au point critique.

Le plus utilisé est le test de déviance.

Le test de la déviance (ou du rapport de vraisemblance)

Pour tester $H_0 : R\beta = 0$ contre $H_1 : R\beta \neq 0$, le principe du test est le suivant :

- On calcule l'EMV dans chaque modèle pour obtenir $\widehat{\beta}$ dans le modèle complet et $\widehat{\beta_{H_0}}$ dans le modèle contraint.
- Si H_0 est vraie, le modèle contraint devrait être aussi "vraisemblable" que le modèle complet, donc $L(\widehat{\beta})$ et $L(\widehat{\beta_{H_0}})$ devraient être similaires.
- La statistique de test est ainsi la différence des déviances :

$$D_{H_0} - D_{H_1} = 2\left(L(\widehat{\beta}) - L(\widehat{\beta_{H_0}})\right)$$

- Sous H_0, en notant q le nb de contraintes, on a la convergence (admise) :

$$D_{H_0} - D_{H_1} = 2\left(L(\widehat{\beta}) - L(\widehat{\beta_{H_0}})\right) \xrightarrow{\mathcal{L}} \chi^2_q$$

- La région critique au niveau asymptotique α est donc

$$RC_\alpha = \left\{D_{H_0} - D_{H_1} > \chi^2_q(1-\alpha)\right\},$$

où $\chi^2_q(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'une loi χ^2_q.

- La p-value vaut

$$p - value = 1 - F_q\left(D_{H_0} - D_{H_1}\right)$$

où F_q désigne la fdr d'une loi χ^2_q.

Cas particulier 1 : test de significativité

- On souhaite tester si un modèle (ayant une constante) est significatif
- On teste donc H_0 : tous ses coef sont nuls sauf la constante.
- Cela correspond au cas particulier $R = \left(0 \,|\, I_{p-1}\right)$.
- Il suffit de comparer la déviance du modèle à la déviance nulle D_0, correspondant à un modèle qui ne contient que la constante.
- La statistique de test est $D_0 - D$. Sous H_0, lorsque $n \to \infty$:

$$D_0 - D \sim \chi^2_{p-1}$$

- Le modèle est donc significatif (par rapport au modèle nul) si l'échantillon est dans la région critique de niveau asymptotique α :

$$RC_\alpha = \{D_0 - D > \chi^2_{p-1}(1-\alpha)\}.$$

Cas particulier 2 : test de comparaison de modèles emboités

Si deux modèles sont emboités, on peut de même les comparer en comparant leur déviance.

- Supposons que le modèle 1 (de déviance D_1) soit un sous-modèle du modèle 2 (de déviance D_2)
- Le modèle 1 est donc obtenu à partir du modèle 2, de paramètre β, via une contrainte du type $R\beta = 0$ où R est une matrice (q, p).
- Sous $H_0 : R\beta = 0$, on a asymptotiquement $D_1 - D_2 \sim \chi^2_q$
- D'où le test asymptotique : $RC_\alpha = \{D_1 - D_2 > \chi^2_q(1-\alpha)\}$.
- Intuitivement, le modèle le plus simple est préférable lorsque la déviance n'a pas décru suffisamment en passant au plus gros modèle.
- Le test de significativité précédent est le cas particulier où $D_1 = D_0$

Les critères AIC et BIC sont définis de façon similaire à la régression linéaire, c'est à dire

$$AIC = -2L_{mod} + 2p, \qquad BIC = -2L_{mod} + \ln(n)p.$$

où L_{mod} est la log-vraisemblance du modèle estimé.

Cela revient (quitte à oublier L_{sat} qui agit comme une constante) à :

$$AIC = D + 2p, \qquad BIC = D + \ln(n)p.$$

En pratique :

- On choisit le modèle ayant l'AIC ou le BIC minimal
- Comme en régression linéaire, on peut utiliser des procédures de sélection automatique (backward, forward, etc).

Exemple : étude sur l'obésité aux Etats-Unis

Lorsqu'on estime le modèle

```
glm(Y~AGE+DBP+SEXE+ACTIV+WALK+MARITAL, family=binomial)
```

on obtient dans le `summary` :

```
    Null deviance: 4610.8  on 5300  degrees of freedom
Residual deviance: 4459.5  on 5290  degrees of freedom
  AIC: 4481.5
```

La déviance du modèle vaut donc $D = 4459.5$.

Test de significativité : on compare D à la déviance nulle $D_0 = 4610.8$.
$D_0 - D = 151.3$. La p-value du test vaut $1 - F_{10}(151.3) \approx 0$ où F_{10} est la fdr d'une χ^2_{10}. Le modèle est significatif.

Exemple : étude sur l'obésité aux Etats-Unis

On veut tester si la variable MARITAL est utile ou non. On estime

```
glm(Y∼AGE+DBP+SEXE+ACTIV+WALK, family=binomial)
```

et on obtient :

```
    Null deviance: 4610.8  on 5300  degrees of freedom
Residual deviance: 4462.7  on 5295  degrees of freedom
AIC: 4474.7
```

La déviance vaut cette fois-ci $D_2 = 4462.7$.

On vérifie que le modèle est bien significatif.

Pour comparer au modèle précédent, on calcule : $D_2 - D = 3.2$.
La p-value du test vaut $1 - F_5(3.2) \approx 0.67$, où F_5 : fdr d'une χ^2_5.
On accepte donc H_0 : les coefficents liés à MARITAL sont nuls.

Cela se confirme aussi via l'AIC (qui est plus faible ici).

Pour le modèle avec AGE^2 :

```
glm(Y~AGE+I(AGE^2)+DBP+SEXE+WALK+ACTIV, family=binomial)
```

on obtient :

```
    Null deviance: 4610.8  on 5300  degrees of freedom
Residual deviance: 4439.5  on 5294  degrees of freedom
AIC: 4453.5
```

- On vérifie que ce modèle est bien significatif.
- La test de deviance avec le modèle précédent a pour p-value :
  $1 - F_1(4462.7 - 4439.5) = 1 - F_1(23.2) \approx 10^{-6}$.
  Ce modèle est donc préférable, ce qui se confirme via les AIC.
- Par contre, on ne peut pas comparer ce modèle avec le premier par le test de deviance car ils ne sont pas emboités.

  On peut cependant comparer leur AIC : ce modèle est préférable.

Supposons qu'on s'intéresse à un nouvel individu pour lequel

- on connait ses caractéristiques $x \in \mathbb{R}^p$,
- on ne connait pas son Y.

On souhaite prédire Y pour ce nouvel individu.

Si on a ajusté un modèle de régression logistique, on peut estimer $p_\beta(x) = \mathbb{P}(Y = 1|X = x)$ par

$$p_{\hat{\beta}}(x) = logit^{-1}(x'\hat{\beta}) = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}.$$

On va voir :

1. comment construire un intervalle de confiance autour de cette estimation ;
2. comment exploiter cette estimation pour classer l'individu dans la catégorie $Y = 0$ ou $Y = 1$.

IC pour $p_\beta(x)$

- On sait que lorsque $n \to \infty$: $\hat{\beta} \sim \mathcal{N}(\beta, (X' W_{\hat{\beta}} X)^{-1})$.
- Ainsi lorsque $n \to \infty$: $x'\hat{\beta} \sim \mathcal{N}(x'\beta, x'(X' W_{\hat{\beta}} X)^{-1} x)$.
- On en déduit un IC asymptotique pour $x'\beta$

$$IC_{1-\alpha}(x'\beta) = \left[x'\hat{\beta} \pm q(1-\alpha/2) \sqrt{x'(X' W_{\hat{\beta}} X)^{-1} x} \right].$$

- Puisque $p_{\hat{\beta}}(x) = logit^{-1}(x'\hat{\beta})$, on a donc par application de la fonction croissante $logit^{-1}$, l'IC au niveau asymptotique $1-\alpha$:

$$IC_{1-\alpha}(p_\beta(x)) = \left[logit^{-1}\left(x'\hat{\beta} - q(1-\alpha/2) \sqrt{x'(X' W_{\hat{\beta}} X)^{-1} x} \right) ; \right.$$
$$\left. logit^{-1}\left(x'\hat{\beta} + q(1-\alpha/2) \sqrt{x'(X' W_{\hat{\beta}} X)^{-1} x} \right) \right].$$

On a estimé $p_\beta(x) = \mathbb{P}(Y = 1|X = x)$ par $p_{\hat{\beta}}(x)$.

Pour un seuil **à choisir** $s \in [0, 1]$, on utilise la règle :

$$\begin{cases} \text{si} & p_{\hat{\beta}}(x) > s, \quad \hat{Y} = 1, \\ \text{si} & p_{\hat{\beta}}(x) < s, \quad \hat{Y} = 0. \end{cases}$$

Le choix "naturel" du seuil est $s = 0.5$ mais ce choix peut être optimisé.

On procède par validation croisée : on estime Y sur un échantillon test et on forme la matrice de confusion.

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	VN	FP
$Y=1$	FN	VP

- VN : nombre de *vrais négatifs* : nombre d'individus ayant été classés négatifs ($\hat{Y}=0$), et étant réellement négatifs ($Y=0$)
- FN : nombre de *faux négatifs* : nombre d'individus ayant été classés négatifs ($\hat{Y}=0$), et étant en fait positifs ($Y=1$)
- FP : nombre de *faux positifs* : nombre d'individus ayant été classés positifs ($\hat{Y}=1$), et étant en fait négatifs ($Y=0$)
- VP : nombre de *vrais positifs* : nombre d'individus ayant été classés positifs ($\hat{Y}=1$), et étant réellement positifs ($Y=1$)

L'idéal est d'avoir une matrice de confusion la plus diagonale possible.

On cherche généralement à maximiser les indicateurs suivants :

- La sensibilité (ou rappel, ou recall, ou taux de vrais positifs) estime $\mathbb{P}(\hat{Y} = 1 | Y = 1)$ par $VP/(VP + FN)$.
- La spécificité (ou sélectivité, ou taux de vrais négatifs) estime $\mathbb{P}(\hat{Y} = 0 | Y = 0)$ par $VN/(VN + FP)$.
- La précision (ou valeur prédictive positive) estime $\mathbb{P}(Y = 1 | \hat{Y} = 1)$ par $VP/(VP + FP)$.

ou encore le F-score qui est la moyenne harmonique entre sensibilité et précision :

$$F_1 = 2 \frac{\textit{précision} \times \textit{sensibilité}}{\textit{précision} + \textit{sensibilité}}.$$

Choix du seuil s

Pour chaque seuil s, à partir d'un échantillon test :

- on peut former la matrice de confusion
- calculer des scores (sensibilité, F-score, etc)

On choisit finalement le seuil s optimal, selon le score d'intérêt.

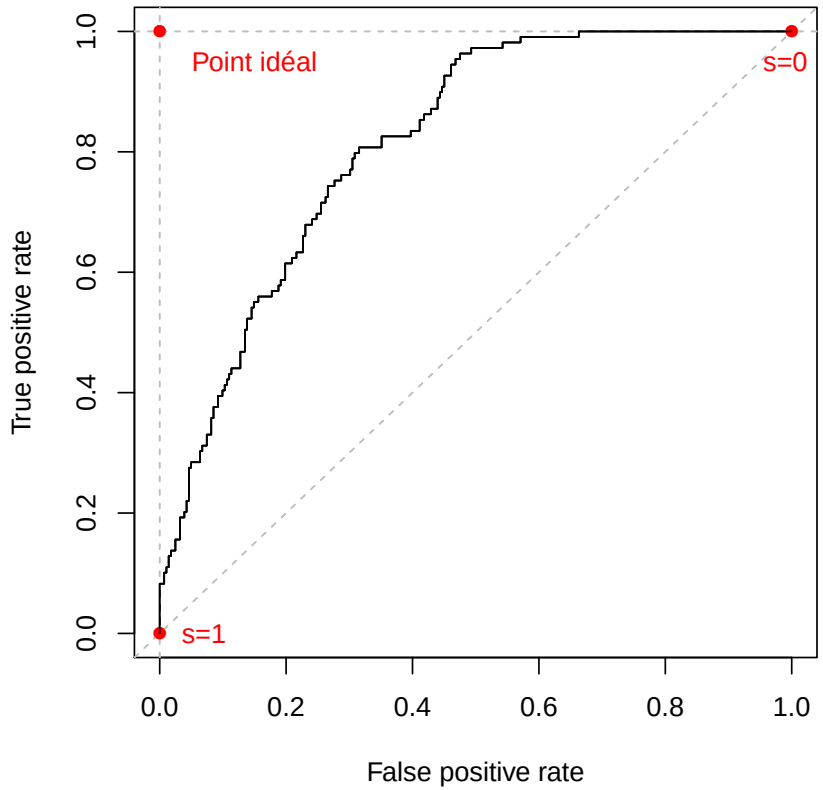
Le score d'intérêt...

- il dépend du contexte de l'étude
- il peut être beaucoup plus grave de prédire à tort $\hat{Y} = 0$ que $\hat{Y} = 1$
  Exemple : $Y = 1$ si le patient présente une maladie grave
- ou l'inverse...

Courbe ROC

On peut également tracer la courbe ROC (taux de vrais positifs en fonction du taux de faux positifs pour $s \in [0, 1]$) :

- Elle résume le pouvoir discriminant du modèle pour tout s.
- L'AUC (aire sous la courbe) est un indicateur de qualité du modèle ($0 \leq AUC \leq 1$). Ou équivalent, l'indice de Gini : $2 \times AUC - 1$.
- La courbe ROC est utile pour se comparer à un autre modèle de classification.



4 Modèles linéaires généralisés

- Généralité sur les GLM (generalized linear models)
- Le modèle logistique pour Y binaire
- Modèles pour données catégorielles
  - Modèle logistique nominal
  - Modèle logistique ordinal
- Modèles pour données de comptage

- Soit Y une variable qualitative qui peut prendre $K \geq 2$ modalités.
- Sans perte de généralité, $Y \in \{0, \ldots, K-1\}$.
- Soit $X \in \mathbb{R}^p$ un vecteur de régresseur. On cherche à modéliser

$$p^{(k)}(x) = \mathbb{P}(Y = k | X = x),$$

  pour tout $k \in \{0, \ldots, K-1\}$ et $x \in \mathbb{R}^p$.
- Le cas $K = 2$ revient à la partie précédente.

Deux situations seront distinguées :

- Cas nominal : aucun ordre possible entre les modalités
  Exemple : CSP, Gouts musicaux,...
- Cas ordinal : les modalités sont naturellement ordonnées
  Exemple : niveau de satisfaction, niveau de rémission,...

- Dans le cas binaire ($K = 2$), le modèle logistique suppose qu'il existe $\beta \in \mathbb{R}^p$ tel que :

$$\frac{p^{(1)}(x)}{p^{(0)}(x)} = e^{x'\beta}.$$

  La modalité "0" peut être vue comme une modalité de référence.

- Dans le cas général (K quelconque), le modèle logistique nominal (ou multinomial, ou à modalité de référence) suppose de même :

$$\frac{p^{(k)}(x)}{p^{(0)}(x)} = e^{x'\beta^{(k)}}, \quad \text{pour } k \in \{1, \dots, K-1\},$$

  où $\beta^{(k)} \in \mathbb{R}^p$ est la paramètre associé à la modalité k.

- La modalité "0" est la modalité de référence, dont la probabilité se déduit des autres.

- Il y a en tout $(K-1) \times p$ paramètres inconnus.

On en déduit que dans ce modèle, pour tout $k \in \{1, \dots, K-1\}$,

$$p^{(k)}(x) = p_\beta^{(k)}(x) = \frac{e^{x'\beta^{(k)}}}{1 + \sum_{r=1}^{K-1} e^{x'\beta^{(r)}}},$$

tandis que

$$p^{(0)}(x) = p_\beta^{(0)}(x) = \frac{1}{1 + \sum_{r=1}^{K-1} e^{x'\beta^{(r)}}}.$$

(qui est cohérent avec la formule précédente en prenant $\beta^{(0)} = 0$)

- On remarque que chaque $p_\beta^{(k)}(x)$ dépend bien de tous les paramètres $\beta = (\beta^{(1)}, \dots, \beta^{(K-1)})$ et non pas seulement de $\beta^{(k)}$
- D'où la notation $p_\beta^{(k)}(x)$ avec l'indice β.

- La valeur des paramètres $\beta^{(k)}$ dépend de la modalité de référence.
- On appelle "cote" de l'événement $Y = k$, le rapport $p_\beta^{(k)}(x)/p_\beta^{(0)}(x)$.
- L'OR de $Y = k$ pour deux caractéristiques x_1 et x_2 est donc

$$OR^{(k)}(x_1, x_2) = \frac{p_\beta^{(k)}(x_1)/p_\beta^{(0)}(x_1)}{p_\beta^{(k)}(x_2)/p_\beta^{(0)}(x_2)} = e^{(x_1 - x_2)' \beta^{(k)}}$$

- Il ne dépend que de $\beta^{(k)}$, et même, que de $\beta_j^{(k)}$ si x_1 et x_2 ne diffère que par le régresseur $X^{(j)}$.
- On retrouve la même interprétation des OR qu'en régression logistique, sauf qu'ici **la cote est relative à la modalité de référence**.
- Il est donc important de choisir judicieusement la modalité de référence pour les interprétations.

- Ceci dit, pour deux modalités $k \neq l$, le rapport de probabilité

$$\frac{p_\beta^{(k)}(x)}{p_\beta^{(l)}(x)} = e^{x'(\beta^{(k)}-\beta^{(l)})}$$

  ne dépend pas de la modalité de référence choisie.

- De même, la valeur des probabilités $p_\beta^{(k)}(x)$ et leur estimation ne dépendent pas de la modalité de référence choisie.

- En fait, si la modalité de référence est $Y = j$, en notant les paramètres associés $\gamma^{(k)}$, $k \neq j$, et $\gamma^{(j)} = 0$, on a la relation $\gamma^{(k)} = \beta^{(k)} - \beta^{(j)}$, ce qui justifie les deux affirmations précédentes.

- La variable aléatoire $Y|X = x$ suit une loi multinomiale (associée à un seul "lancer") de paramètres $p_\beta^{(0)}(x), \ldots, p_\beta^{(K-1)}(x)$.
- La vraisemblance d'un échantillon $(Y_1|X = x_1), \ldots, (Y_n|X = x_n)$ s'écrit (voir tableau) :

$$\prod_{i=1}^{n} \prod_{k=0}^{K-1} \left(p_\beta^{(k)}(x_i) \right)^{\mathbf{1}_{Y_i=k}}$$

- Donc la log-vraisemblance

$$L = \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbf{1}_{Y_i=k} \ln \left(p_\beta^{(k)}(x_i) \right)$$

- Dans le cas du modèle logistique nominal, on en déduit (cf tableau)

$$L = \sum_{i=1}^{n} \left(\sum_{k=1}^{K-1} x_i' \beta^{(k)} \mathbf{1}_{y_i=k} - \ln \left(1 + \sum_{k=1}^{K-1} e^{x_i' \beta^{(k)}} \right) \right)$$

- En annulant le gradient de L, on obtient que $\hat{\beta} = (\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(K-1)})$ doit vérifier

$$\sum_{i=1}^{n} x_i \mathbf{1}_{y_i=k} = \sum_{i=1}^{n} x_i p_{\hat{\beta}}^{(k)}(x_i), \quad \text{pour tout } k \in \{1, \ldots, K-1\}.$$

- Il s'agit de $K-1$ systèmes, ayant chacun p équations (car $x_i \in \mathbb{R}^p$),
- soit un système à $(K-1) \times p$ équations.
- On le résout numériquement pour trouver $\hat{\beta}$ de taille $(K-1) \times p$.

Comme en régression logistique :

- On peut montrer que L est strictement concave dès lors que $rg(\mathbf{X}) = p$. Cela assure l'unicité de l'EMV (s'il existe).
- L'existence est assurée si aucune modalité n'est séparée des autres par un hyperplan.
- Sous des hypothèses de régularité semblables au cas de la régression logistique, on a

$$J_n(\beta)^{1/2}\left(\widehat{\beta} - \beta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_{(K-1)\times p})$$

où $J_n(\beta)$ est la matrice d'information de Fisher et $I_{(K-1)\times p}$ la matrice identité de taille $(K-1) \times p$.

- La matrice $J_n(\beta)$, non détaillée ici, est une matrice de $(K-1) \times (K-1)$ blocs, chacun ayant une forme similaire à la matrice d'information de Fisher de la régression logistique.

Les outils d'inférence sont basés sur la loi asymptotique de $\hat{\beta}$ et sont similaires à ceux de la régression logistique :

- La significativité de chaque coefficient peut-être testé par un test (asymptotique) de Wald.
- Des intervalles de confiance, pour les coefficients et les OR, s'en déduisent de façon analogue.

La déviance se définit de manière similaire : $D = 2(L_{sat} - L_{mod})$.
Comme on le fait pour la régression logisitique :

- On peut tester la significativité du modèle,
- ou comparer deux modèles emboités.

Puisqu'on estime $p(K-1)$ paramètres, on a enfin :

$$AIC = D + 2\,p(K-1), \qquad BIC = D + \ln(n)\,p(K-1).$$

Exemple : Etude sur l'équipement des voitures

Préférence pour une voiture équipée (avec climatisation et direction assistée), selon la classe d'âge et le sexe.

Genre	Catégorie d'âge	Pas important	Important	Très important
Femme	$18-23$	26	12	7
	$24-40$	9	21	15
	>40	5	14	41
Homme	$18-23$	40	17	8
	$24-40$	17	15	12
	>40	8	15	18

- On veut modéliser la variable Y ="importance" (3 modalités)
- Les régresseurs sont le sexe (2 classes) et l'âge (3 classes).
- Il s'agit d'observations répétées : chaque caractéristique croisée âge/sexe est observée chez plusieurs individus.

Exemple : Etude sur l'équipement des voitures

Sous R, on peut utiliser la fonction `vglm` de la libraire `VGAM` :

```
vglm( Y~ age + sexe, family=multinomial)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.5908     0.2840  -2.080 0.037484 *
(Intercept):2  -1.0391     0.3305  -3.144 0.001667 **
age2:1          1.1283     0.3416   3.302 0.000958 ***
age2:2          1.4781     0.4009   3.687 0.000227 ***
age3:1          1.5877     0.4029   3.941 8.12e-05 ***
age3:2          2.9168     0.4229   6.897 5.32e-12 ***
sexeM:1        -0.3881     0.3005  -1.292 0.196510
sexeM:2        -0.8130     0.3210  -2.532 0.011326 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])
```

Remarques :
Dans la sortie Y est encodé 1, 2 ou 3 (de "Pas important" à "très important")
La modalité de référence pour Y est $Y = 1$ ("Pas important")
Age est encodé 1, 2 ou 3 (en croissant)
Les deux modèles estimés sont numérotés 1 et 2.

Exemple : Etude sur l'équipement des voitures

- Les régresseurs sont age (3 modalités) et sexe (2 modalités), et la constante, ce qui fait $(3-1)+(2-1)+1=4$ paramètres à estimer par modalité estimé de Y.
- Y a 3 modalités (1, 2, 3), donc on estime 2 modèles : $p^{(2)}(x)/p^{(1)}(x)$ et $p^{(3)}(x)/p^{(1)}(x)$, la première modalité étant la référence.
- Il y a donc en tout 8 paramètres estimés (4 par modèle).
- Par exemple, pour une femme âgée de 18 à 23 ans :

$$\frac{\mathbb{P}(Y=\text{“}Important\text{”}|Femme\ 18-23)}{\mathbb{P}(Y=\text{“}Pas\ important\text{”}|Femme\ 18-23)}=exp(-0.59)\approx 0.55$$

  Pour un homme de la même classe d'âge, ce rapport vaut $exp(-0.59-0.3881)\approx 0.38$.
- L'OR entre un homme et une femme pour la préférence "Très importante" par rapport à "Pas important" vaut $exp(-0.813)=0.44$.
  Cette cote est donc plus que double chez les femmes que chez les hommes, toutes choses égales par ailleurs.

Si les modalités de Y suivent un ordre naturel :

- On peut évidemment l'ignorer et utiliser le modèle nominal précédent : il est très général mais a beaucoup de paramètres.
- Mais on peut tirer partie de cette structure pour simplifier le modèle (moins de paramètres, interprétation plus aisée).

Pour un modèle nominal :

- En cohérence avec le modèle logistique, on a focalisé sur la "cote"

$$\frac{p^{(k)}(x)}{p^{(0)}(x)} = \frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = 0|X = x)},$$

- de telle sorte que $\beta^{(k)}$ s'interprète facilement via $OR^{(k)}(x_1, x_2)$.
- Cet OR quantifie à quel point la "cote" de $\mathbb{P}(Y = k)$ est modifiée entre x_1 et x_2, relativement à la modalité de référence $Y = 0$.
- Dans le cas ordinal, on va modéliser des "cotes" plus faciles à interpréter.

Quelle cote modéliser lorsque les modalités sont ordonnées ?

Pour un modèle ordinal, plusieurs alternatives sont envisageables :

- (Garder le point de vue nominal)
- Le modèle à catégories adjacentes modélise les "cotes" suivantes :

$$\frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y=k-1|X=x)}.$$

- Le modèle logistique "continuous ratio" modélise les "cotes" suivantes :

$$\frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y\leq k-1|X=x)}.$$

- Le modèle à odds proportionnels modélise les cotes suivantes :

$$\frac{\mathbb{P}(Y\leq k|X=x)}{\mathbb{P}(Y>k|X=x)}.$$

  (Il s'agit pour une fois d'une vraie cote, de la forme $p/(1-p)$)
- Dans chaque cas, les paramètres auront une interprétation naturelle pour l'OR correspondant.

Le modèle à odds proportionnels est le plus utilisé en pratique.

L'idée est de construire des modèles logistiques pour les variables binaires $\mathbf{1}_{Y \leq k}$, pour tout $k \in \{0, \dots, K-2\}$.

- Cela donne en toute généralité le **modèle cumulatif**

$$logit(\mathbb{P}(Y \leq k | X = x)) = \ln \frac{\mathbb{P}(Y \leq k | X = x)}{\mathbb{P}(Y > k | X = x)} = x' \beta^{(k)}, \quad k \in \{0, \dots, K-2\}.$$

  Ce modèle a $p(K-1)$ paramètres mais est différent du modèle nominal.

- Le modèle à **odds proportionnels** suppose que l'effet des régresseurs (hormis la constante) est constant quelles que soient les modalités :

$$logit(\mathbb{P}(Y \leq k | X = x)) = \beta_0^{(k)} + \beta' X^*, \quad k \in \{0, \dots, K-2\},$$

  où $X^* \in \mathbb{R}^{p-1}$ désigne le vecteur des régresseurs autres que la constante.

  Il contient $(K-1) + (p-1)$ paramètres (c'est beaucoup moins).

Modèle logistique à odds proportionnels : existence

Puisque quel que soit k,

$$logit(\mathbb{P}(Y \leq k|X = x)) \leq logit(\mathbb{P}(Y \leq k+1|X = x)),$$

le modèle à odds proportionnel doit vérifier :

$$\beta_0^{(0)} \leq \cdots \leq \beta_0^{(K-2)}.$$

Cette contrainte est imposée lors de l'estimation.

De même pour le modèle cumulatif, on doit avoir *pour tout* x :

$$x'\beta^{(0)} \leq \cdots \leq x'\beta^{(K-2)}$$

- En toute rigueur, cela est impossible (les hyperplans pour deux vecteurs normaux différents se croisent forcément)
- Mais cela peut être ok pour les plages de valeurs observées de x.
- Dans ce cas, le modèle cumulatif est acceptable "localement".
- A défaut, le modèle cumulatif peut ne pas être possible (l'estimation échoue).

Motivation du modèle à odds proportionnels

Supposons que les classes $Y = k$ proviennent de la discrétisation d'une variable latente continue Z : pour $\alpha_{-1} = -\infty$, $\alpha_0 < \cdots < \alpha_{K-1}$ et $k \in \{0, \ldots, K-1\}$,

$$Y = k \quad \Longleftrightarrow \quad \alpha_{k-1} \leq Z < \alpha_k.$$

Exemple : Z est une note, et Y est une mention.

Supposons qu'il existe un lien linéaire entre Z et les régresseurs X :

$$Z = \beta' X + \varepsilon$$

où ε suit une loi de fdr F.
Alors

$$\mathbb{P}(Y \leq k) = F(\alpha_k - \beta' X).$$

- La dépendance en X ne dépend pas de k.
- Si $F = logit^{-1}$, on obtient le modèle à odds proportionnels.
- D'autres choix de F sont possibles (probit,...) mais les OR deviennent moins interprétables.

Pourquoi "odds proportionnels"?

Le modèle à odds proportionnel suppose que

$$\frac{\mathbb{P}(Y \leq k|X = x)}{\mathbb{P}(Y > k|X = x)} = e^{\beta_0^{(k)} + \beta' x^*}, \quad k \in \{0, \ldots, K-2\}.$$

Il s'agit de la cote (l'odds) de $Y \leq k$ sachant X.

L'odds-ratio de $Y \leq k$ entre deux individus de régresseurs x_1 et x_2 respectivement vaut donc

$$OR(x_1, x_2) = \exp\left(\beta'(x_1^* - x_2^*)\right).$$

- Cet OR ne dépend pas de k.
- $\ln(OR(x_1, x_2))$ est "proportionnel" à $(x_1^* - x_2^*)$, la "constante" de proportionnalité β (en fait un vecteur) étant indépendante de k.

Égalité des pentes

L'écriture

$$logit(\mathbb{P}(Y \leq k|X)) = \beta_0^{(k)} + \beta' X^*, \quad k \in \{0, \ldots, K-2\},$$

implique que les $K-1$ hyperplans

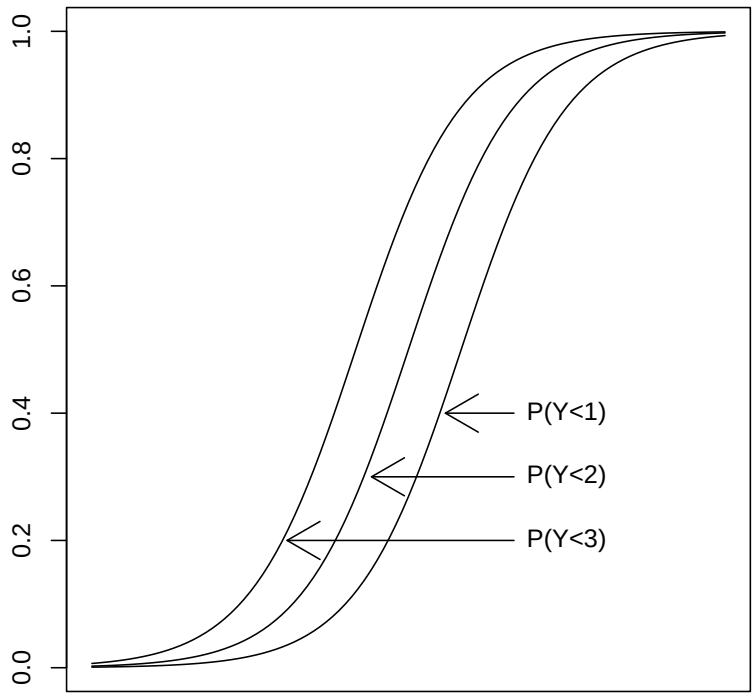
$$X^* \mapsto logit(\mathbb{P}(Y \leq k|X)), \quad k \in \{0, \ldots, K-2\},$$

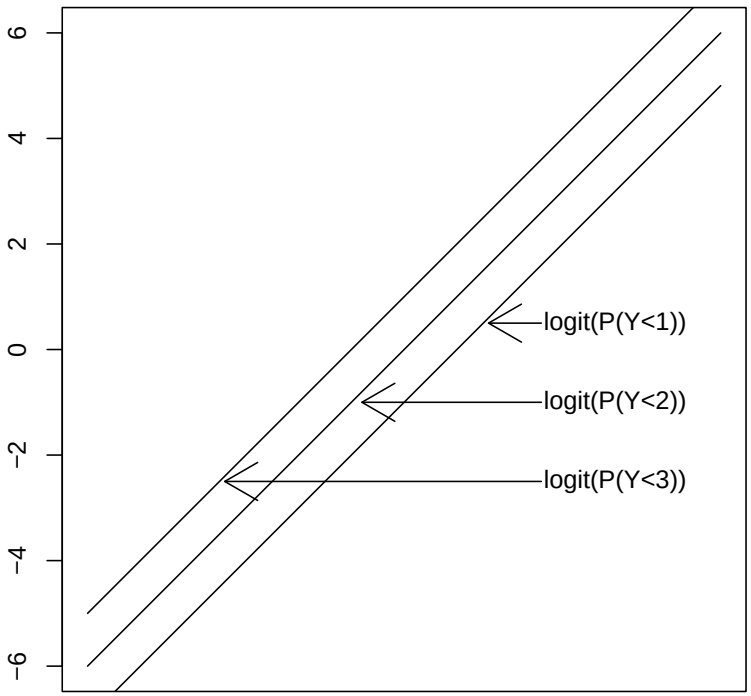
sont parallèles.
Ils ont en effet tous les même vecteur normal β.
Ils ne diffèrent que par la constante à l'origine $\beta_0^{(k)}$.

Pour valider le modèle à odds proportionnel, il convient de tester si cette propriété est vraie.

Égalité des pentes : illustration graphique

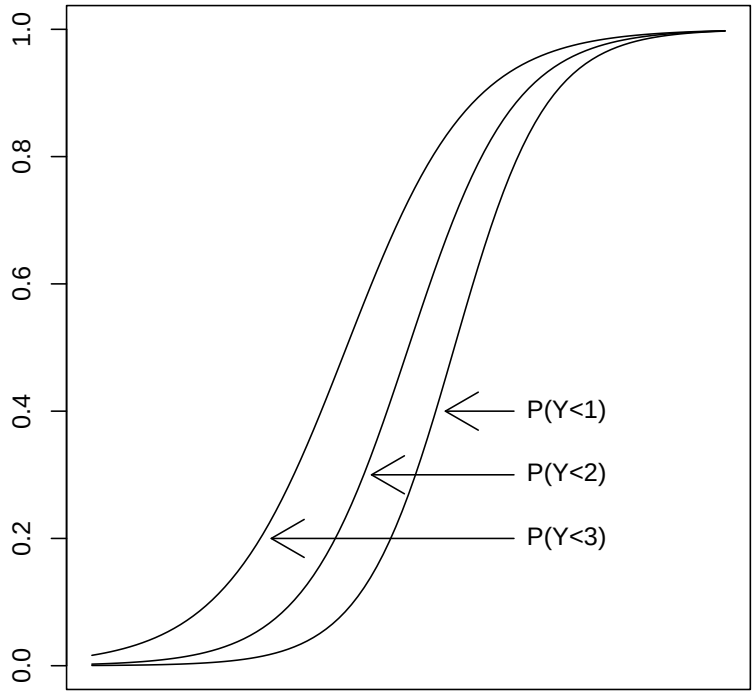
Exemple d'égalité des pentes (cas où X^* est de dimension 1)

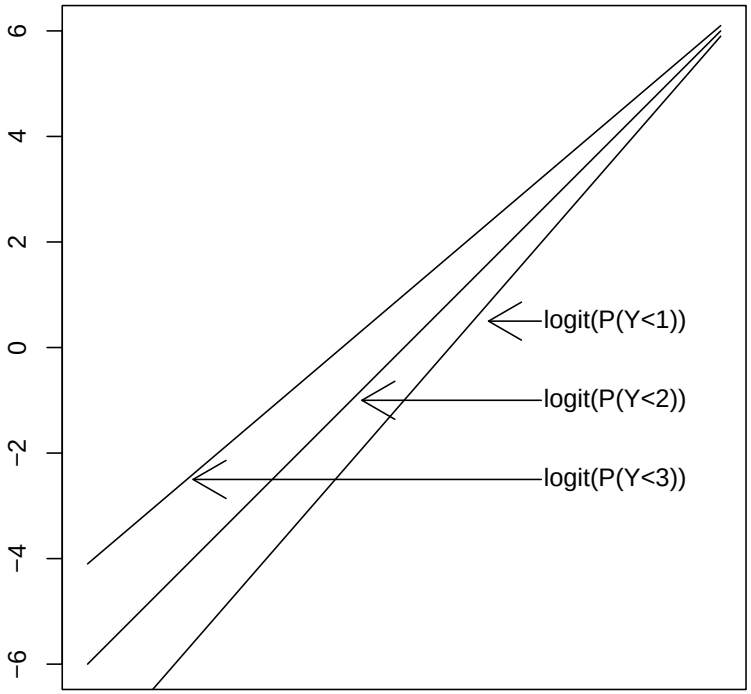




Égalité des pentes : illustration graphique

Exemple d'absence d'égalité des pentes (cas où X^* est de dimension 1)





Pour tester l'égalité des pentes, il suffit de partir du modèle cumulatif général

$$\frac{\mathbb{P}(Y \leq k|X=x)}{\mathbb{P}(Y > k|X=x)} = e^{x'\beta^{(k)}}, \quad k \in \{0,\dots,K-2\}$$

et tester si les paramètres (sauf la constante) sont égaux quel que soit k. En écrivant $\beta^{(k)} = (\beta_0^{(k)},\dots,\beta_{p-1}^{(k)})$, $\beta_0^{(k)}$ étant la constante, on teste :

$$H_0 : \begin{cases} \beta_1^{(0)} = \cdots = \beta_1^{(K-2)}, \\ \vdots \\ \beta_{p-1}^{(0)} = \cdots = \beta_{p-1}^{(K-2)}. \end{cases}$$

Cela peut se faire par un test de déviance (du rapport de vraisemblance) en comparant le modèle cumulatif général et le modèle à odds proportionnel.
$\rightarrow$ Uniquement si le modèle cumulatif est possible pour le jeu de données.
$\rightarrow$ Sous H_0, la loi asymptotique est une $\chi^2_{(p-1)(K-2)}$.

Rappel : les $(Y_i|X = x_i)$ étant indépendants et de loi multinomiale, la log-vraisemblance s'écrit

$$L = \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbf{1}_{Y_i=k} \ln \left(p_\beta^{(k)}(x_i) \right)$$

Pour le modèle cumulatif et le modèle à odds proportionnels :

- on peut en déduire la forme de $p_\beta^{(k)}$
- on maximise alors L en β pour obtenir $\hat{\beta}$ (par des méthodes numériques)

Pour les tests, comme d'habitude :

- On peut comparer $L_{mod} = L(\hat{\beta})$ avec d'autres modèles emboités pour faire un test du rapport de vraisemblance (c'est à dire de déviance).
- $\hat{\beta} - \beta$ suit asymptotiquement une $\mathcal{N}(0, J_n(\beta)^{-1})$, où $J_n(\beta)$ est l'opposée de la Hessienne de L.
- On peut donc faire des tests de Wald.

Exemple : Etude sur l'équipement des voitures

Genre	Catégorie d'âge	Pas important	Important	Très important
Femme	$18-23$	26	12	7
	$24-40$	9	21	15
	>40	5	14	41
Homme	$18-23$	40	17	8
	$24-40$	17	15	12
	>40	8	15	18

Rappel :

- On veut modéliser la variable Y ="importance" (3 modalités)
- Les régresseurs sont le sexe (2 classes) et l'âge (3 classes).
- Il s'agit d'observations répétées : chaque caractéristique croisée âge/sexe est observée chez plusieurs individus.

On a déjà modélisé Y à l'aide d'un modèle multinomial.
En fait Y est une variable ordinale : on va l'exploiter.

Exemple : Etude sur l'équipement des voitures

On estime un modèle à odds proportionnels :

```
vglm( Y~ age + sexe, family=cumulative(parallel=TRUE))
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.04354    0.23030   0.189   0.8501
(Intercept):2  1.65498    0.25360   6.526 6.76e-11 ***
age2          -1.14710    0.27727  -4.137 3.52e-05 ***
age3          -2.23246    0.29042  -7.687 1.50e-14 ***
sexeM          0.57622    0.22611   2.548   0.0108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 4.5321 on 7 degrees of freedom

Log-likelihood: -25.6671 on 7 degrees of freedom
```

Remarques :

Dans la sortie Y est encodé 1, 2 ou 3 (de "Pas important" à "très important")
Age est encodé 1, 2 ou 3 (en croissant)

Les deux modèles estimés sont numérotés 1 et 2 (seule la constante diffère)

- Il y a en tout $(K-1)+(p-1)=(3-1)+(4-1)=5$ coefficients
- Seule la constante (intercept) change suivant les modalités de Y

Par exemple, pour une femme âgée de 18 à 23 ans

$$\frac{\mathbb{P}(\text{“}Pas\ important\text{”}|Femme\ 18-23)}{\mathbb{P}(\text{“}Important\ ou\ tres\ important\text{”}|Femme\ 18-23)}=e^{0.043}\approx 1.04$$

Pour une femme âgée de plus de 40 ans

$$\frac{\mathbb{P}(\text{“}Pas\ important\text{”}|Femme\ >40)}{\mathbb{P}(\text{“}Important\ ou\ tres\ important\text{”}|Femme\ >40)}=e^{0.043-2.23}\approx 0.11$$

L'OR $e^{0.57622}=1.78$ montre que la cote d'avoir une préférence moindre est 1.78 fois plus importante pour les hommes que pour les femmes.

Exemple : Etude sur l'équipement des voitures

On peut tester l'hypothèse d'égalité des pentes :

- On lance le modèle cumulatif complet

  `vglm( Y~ age + sexe, family=cumulative)`

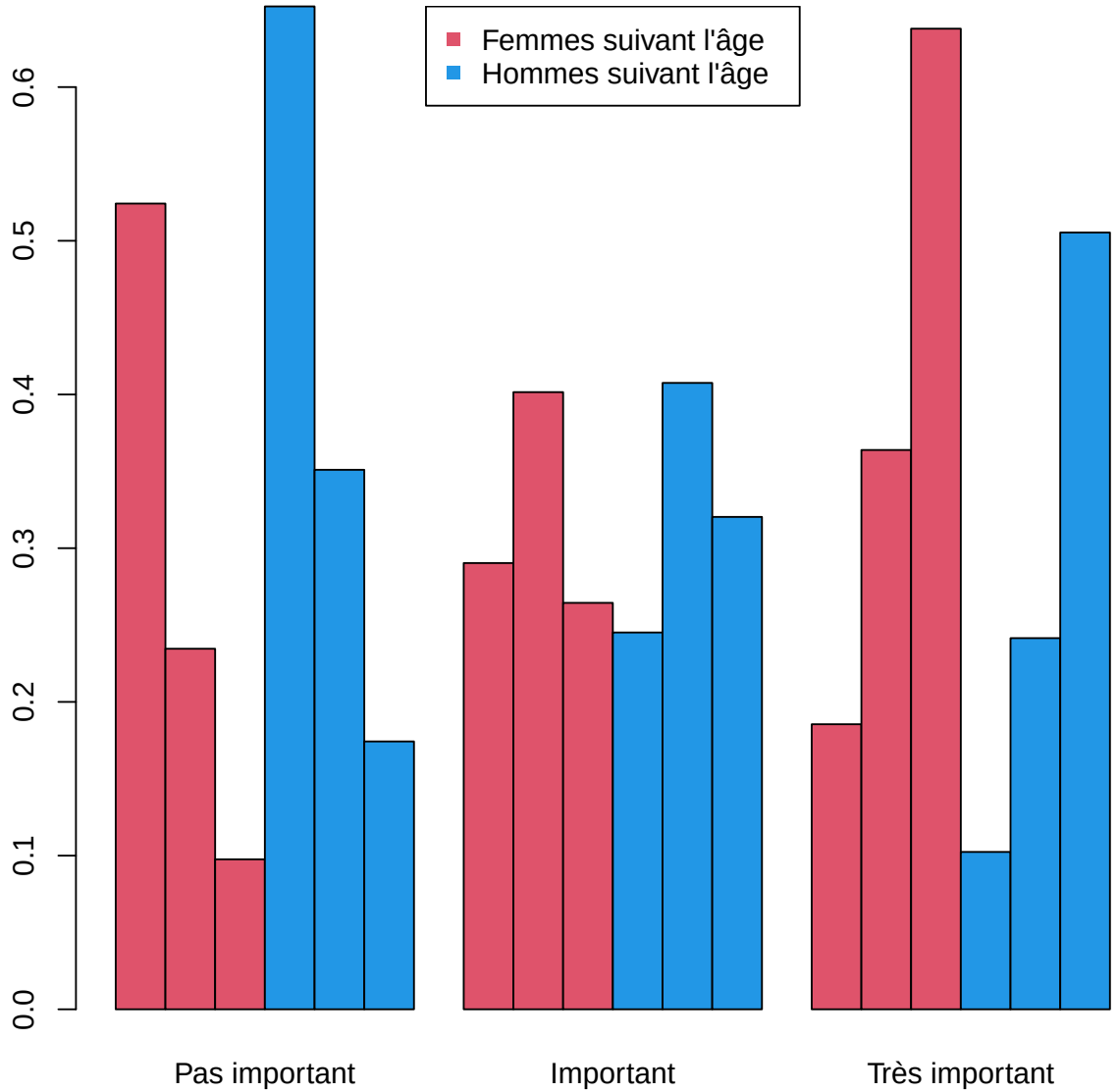
  et on relève la log-vraisemblance qui vaut -25.3164.
- Celle du modèle à odds proportionnel valait -25.6671.
- La stat du test de déviance vaut donc $2 * (25.6671 - 25.3164) = 0.7$.
- On compare à une loi du $\chi^2_{(K-2)(p-1)} = \chi^2_3$: il n'y a pas de raison de rejeter H_0 et donc le modèle à odds proportionnel est préférable au modèle cumulatif.

Pour comparer le modèle à odds proportionnels avec le modèle nominal :

- On ne peut pas utiliser un test de déviance car les deux modèles ne sont pas emboités.
- Néanmoins, l'AIC et le BIC (non reportées ici) sont en faveur du modèle à odds proportionnel (fonctions `AIC` et `BIC` sous `R`).

Exemple : Etude sur l'équipement des voitures

D'après le modèle à odds proportionnels, les probabilités de préférence selon le profil sont estimées ainsi :



4. Modèles linéaires généralisés
   - Généralité sur les GLM (generalized linear models)
   - Le modèle logistique pour Y binaire
   - Modèles pour données catégorielles
   - Modèles pour données de comptage
     - Le modèle log-linéaire de Poisson
     - La sur-dispersion
     - Inflation de zéros

Nous supposons dans ce chapitre que $Y \in \mathbb{N}$

Exemples pour Y : nombre de poissons péchés par jour ; nombre de malades ; nombre de clients ; etc

On dispose de régresseurs $X \in \mathbb{R}^p$ et on cherche à estimer

$$\lambda(x) = \mathbb{E}(Y|X = x).$$

Remarques :

- Contrairement au cas Y binaire ou catégorielle, cette espérance ne revient pas à la probabilité que Y prenne telle ou telle valeur.
- $Y \in \mathbb{N}$ donc $\lambda(x) \geq 0$ (mais $\lambda(x)$ n'est pas forcément un entier).

On va mettre en place un modèle GLM. Pour cela, il faut choisir :

1. la loi de $Y|X$,
2. la fonction de lien g caractérisant la relation, pour $\beta \in \mathbb{R}^p$:

$$\lambda(x) = \mathbb{E}(Y|X = x) = g^{-1}(x'\beta).$$

Fonction de lien. Puisque $\lambda(x) \geq 0$, le choix naturel est $g = \ln$, soit

$$\ln \lambda(x) = x'\beta \quad \Longleftrightarrow \quad \lambda(x) = \exp(x'\beta).$$

Ce choix est rarement remis en cause.

Loi de $Y|X$. Le choix le plus simple, effectué par défaut, est de supposer que $Y|X$ suit une loi de Poisson :

$$Y|X \sim \mathcal{P}(\lambda(X)).$$

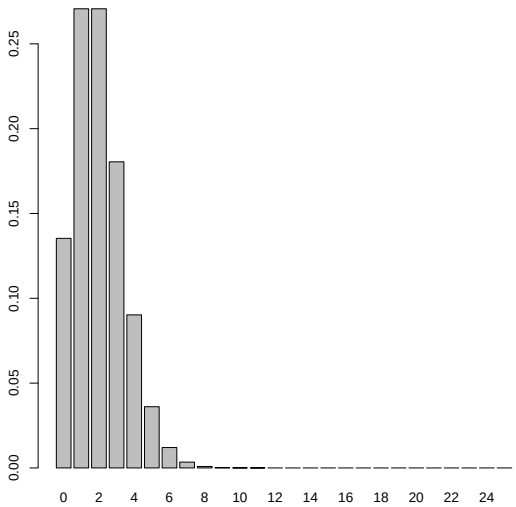
D'autres alternatives sont possibles.

Dans ce chapitre :

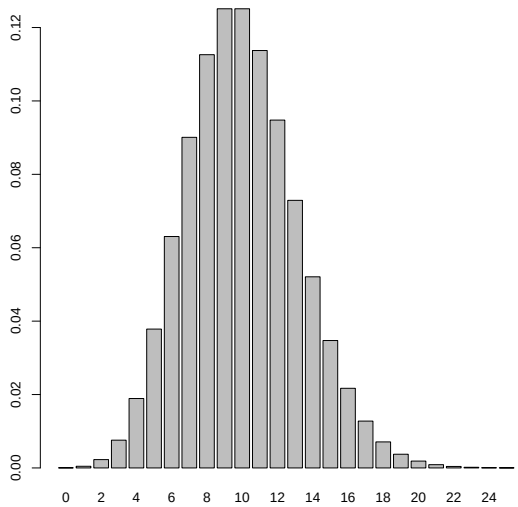
- Nous présentons en détail le cas du modèle log-linéaire de Poisson.
- Nous verrons les alternatives courantes lorsque ce modèle semble inadéquat.

On suppose donc que $Y|X \sim \mathcal{P}(\lambda_\beta(X))$ avec $\lambda_\beta(x) = \exp(x'\beta)$.

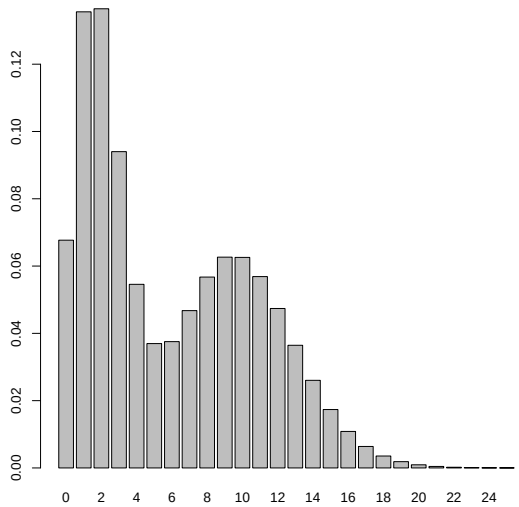
Exemple : régresseur X binaire (avec autant de $X = 0$ que de $X = 1$)



$Y|(X = 0) \sim \mathcal{P}(2)$



$Y|(X = 1) \sim \mathcal{P}(10)$



Y

Pour x fixé, $Y|(X = x) \sim \mathcal{P}(\lambda(x))$ mais Y **ne suit pas une loi de Poisson**.

- Il n'y a pas de notion d'OR puisqu'on n'estime pas une probabilité (ou un ratio de probabilité) mais une espérance.
- La notion équivalente ici est le **rate ratio** (RR), $\lambda(x)$ étant vu comme un taux moyen d'occurence de Y.

Pour deux caractéristiques x_1 et x_2, le rate ratio est simplement :

$$RR(x_1, x_2) = \frac{\lambda_\beta(x_1)}{\lambda_\beta(x_2)} = \exp((x_1 - x_2)'\beta).$$

Si x_1 et x_2 ne diffèrent que par le régresseur j :

$$RR(x_1, x_2) = \exp((x_{1j} - x_{2j})\beta_j).$$

En particulier si ce régresseur est binaire ($x_{1j} = 1$ et $x_{2j} = 0$) : $RR_j = e^{\beta_j}$.

Puisque $Y|(X=x) \sim \mathcal{P}(\lambda_\beta(x))$, on a pour tout $k \in \mathbb{N}$,

$$\mathbb{P}(Y=k|X=x) = e^{-\lambda_\beta(x)} \frac{\lambda_\beta(x)^k}{k!}.$$

Ainsi la vraisemblance de l'échantillon vaut

$$\prod_{i=1}^{n} e^{-\lambda_\beta(x_i)} \frac{\lambda_\beta(x_i)^{y_i}}{y_i!}.$$

Puisque $\lambda_\beta(x) = \exp(x'\beta)$, la log-vraisemblance vaut donc

$$L = \sum_{i=1}^{n} y_i\, x_i'\beta - e^{x_i'\beta} - \ln(y_i!).$$

En annulant le gradient par rapport à β, on trouve que l'EMV $\hat{\beta}$ doit vérifier les équations de vraisemblance

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \lambda_{\hat{\beta}}(x_i) x_i.$$

Il s'agit d'un système à p inconnues qu'on résout numériquement.

Propriétés : comme en régression logistique

- Sous des conditions de régularité, lorsque $n \to \infty$:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X' W_{\hat{\beta}} X)^{-1}),$$

où $W_{\hat{\beta}} = diag(\lambda_{\hat{\beta}}(x_1), \ldots, \lambda_{\hat{\beta}}(x_n))$.

- On peut donc effectuer des tests de significativité de Wald.

Le modèle saturé (un paramètre par observation différente) conduit à

$$\hat{\lambda}(x) = \frac{y_x}{n_x},$$

où

- $y_x = \sum_{i:x_i=x} y_i$ est le nombre total de Y observé pour la caractéristique x sur l'échantillon
- $n_x = \sum_{i:x_i=x} 1$ est le nombre de fois où x a été observé.

La log-vraisemblance du modèle saturé vaut donc :

$$L_{sat} = \sum_x \left(y_x \ln \frac{y_x}{n_x} - y_x \right) - cste$$

où $cste = \sum_{i=1}^n \ln(y_i!)$.

Ainsi la déviance vaut

$$D = 2(L_{sat} - L_{mod}) = 2\sum_x y_x \ln \frac{y_x}{n_x \lambda_{\hat{\beta}}(x)} - (y_x - n_x \lambda_{\hat{\beta}}(x)).$$

Si une constante est dans le modèle (une coordonnée de x vaut 1), on a d'après les équations de vraisemblance $\sum_x y_x = \sum_x n_x \lambda_{\hat{\beta}}(x)$ et alors

$$D = 2\sum_x y_x \ln \frac{y_x}{\hat{y}_x}$$

où $\hat{y}_x = n_x \lambda_{\hat{\beta}}(x)$ sont les effectifs théoriques attendus.

Validation : test de déviance

Comme en régression logistique :

- On peut comparer deux modèles emboités par un test de déviance (ou test du rapport de vraisemblance)
- Si le modèle 2 a q paramètres en moins par rapport au modèle 1, on a sous H_0 : "les q coefficients en question sont nuls" :

$$D_2 - D_1 = 2(L_1 - L_2) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- La région critique de niveau asymptotique α est donc

$$RC_\alpha = \{D_2 - D_1 > \chi^2_q(1-\alpha)\}.$$

- Le test de significativité global correspond au cas où le modèle 2 ne contient que la constante. Dans ce cas $D_2 = D_0$ et $q = p - 1$.

Validation : inspection graphique

Première analyse graphique possible :

- On peut représenter les effectifs prédits $\hat{y}_x = n_x \lambda_{\hat{\beta}}(x)$ par rapport aux effectifs observés y_x.
- Attention : les effectifs prédits $\hat{y}_x$ représentent l'espérance des effectifs attendus sachant x. Il est donc normal que les effectifs observés y_x soient dispersés autour des $\hat{y}_x$.
- Il convient d'avoir des "classes" x suffisamment importantes ($n_x > 5$) pour que le graphique soit pertinent.
- Il est courant de séparer la validation selon certaines catégories d'intérêt (par exemple selon le genre), sous réserve que les effectifs pour chaque "classe croisée" (x/genres) restent suffisants.

Autre analyse graphique : comparer la loi empirique de Y à sa loi prédite.

- La loi empirique de Y est simplement donnée par

$$p_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{y_i=k}, \quad k \in \mathbb{N}.$$

- La loi prédite de Y, pour l'échantillon des x_i observés, est donnée par

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{P}}(Y = k|X = x_i), \quad k \in \mathbb{N},$$

où $\hat{\mathbb{P}}(Y = k|X = x_i)$ est donnée par la loi $\mathcal{P}(\lambda_{\hat{\beta}}(x_i))$, soit

$$\hat{\mathbb{P}}(Y = k|X = x_i) = \frac{\lambda_{\hat{\beta}}(x_i)^k}{k!} e^{-\lambda_{\hat{\beta}}(x_i)} \quad \text{avec } \lambda_{\hat{\beta}}(x_i) = \exp(x_i'\hat{\beta}).$$

Exemple : nombre d'espèces végétales

Nombre d'espèces végétales relevées sur une parcelle en fonction du *pH* du sol (Neutre, Acide ou Basique) et de la biomasse collectée.

Species	pH	Biomass
14	low	3.538
31	mid	0.740
36	high	7.242
20	mid	3.216
⋮	⋮	⋮

On souhaite modéliser $Y =$"Species" en fonction de pH et Biomass.

Exemple : nombre d'espèces végétales

On lance un modèle GLM log-linéaire de Poisson sous R comme ceci :

```
glm(Species ~ pH + Biomass, family=poisson)
```

On obtient :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.84894    0.05281  72.885  < 2e-16 ***
pHlow       -1.13639    0.06720 -16.910  < 2e-16 ***
pHmid       -0.44516    0.05486  -8.114 4.88e-16 ***
Biomass     -0.12756    0.01014 -12.579  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  99.242  on 86  degrees of freedom
AIC: 526.43
```

Exemple : nombre d'espèces végétales

Ainsi, le nombre moyen d'espèces, sachant *pH* et *Biomass*, est

$$\lambda_{\hat{\beta}}(pH, Biomass) = \exp(3.85 - 1.14\,\mathbf{1}_{pH=low} - 0.46\,\mathbf{1}_{pH=mid} - 0.13 Biomass).$$

Le Rate Ratio pour un pH faible (acide) par rapport à un pH élevé (basique) est

$$RR(acide, basique) = \exp(-1.14) = 0.32.$$

En moyenne, il y a donc environ 3 fois moins d'espèces dans un sol acide que dans un sol basique.

On peut essayer d'introduire une interaction entre pH et Biomass

```
glm(Species ~ pH + Biomass + pH:Biomass, family=poisson)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow:Biomass -0.15503    0.04003  -3.873 0.000108 ***
pHmid:Biomass -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39
```

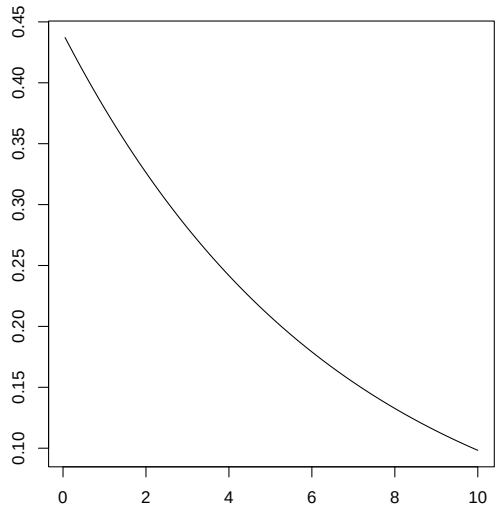
Exemple : nombre d'espèces végétales

- Le modèle avec interaction semble préférable (via AIC et test de deviance).
- Le nombre moyen d'espèces, sachant *pH* et *Biomass*, est cette fois :

$$\lambda_{\hat{\beta}}(pH, Biomass) = \exp(3.77 - 0.82\,\mathbf{1}_{pH=low} - 0.33\,\mathbf{1}_{pH=mid} - 0.11 Biomass - 0.16 Biomass\mathbf{1}_{pH=low} - 0.032 Biomass\mathbf{1}_{pH=mid}).$$

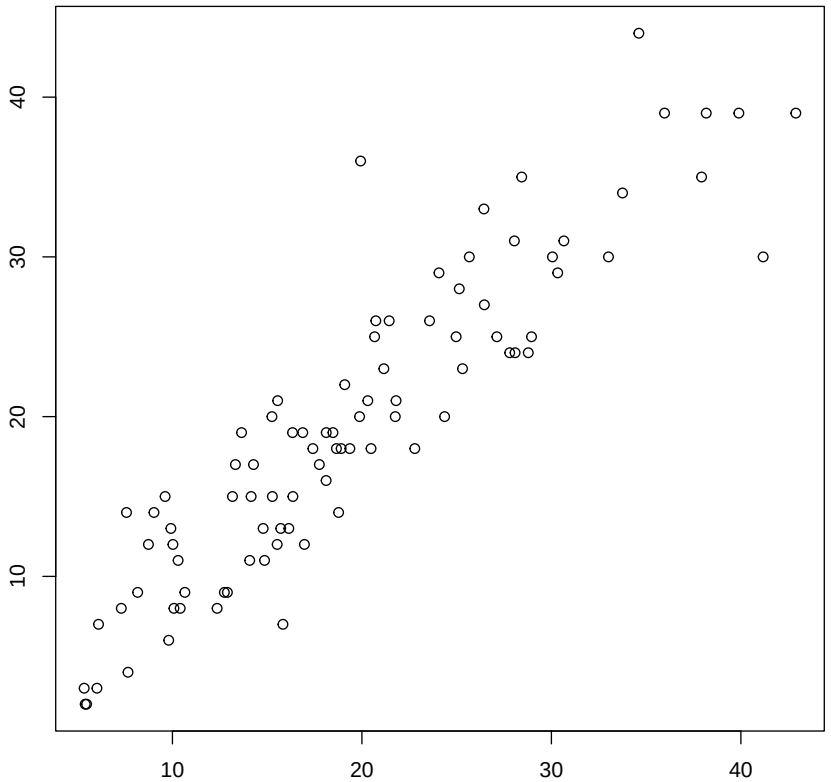
- Le Rate Ratio pour un pH faible (acide) par rapport à un pH élevé (basique) dépend de Biomass et vaut :

$$RR(acide, basique) = \exp(-0.82 - 0.16 Biomass).$$



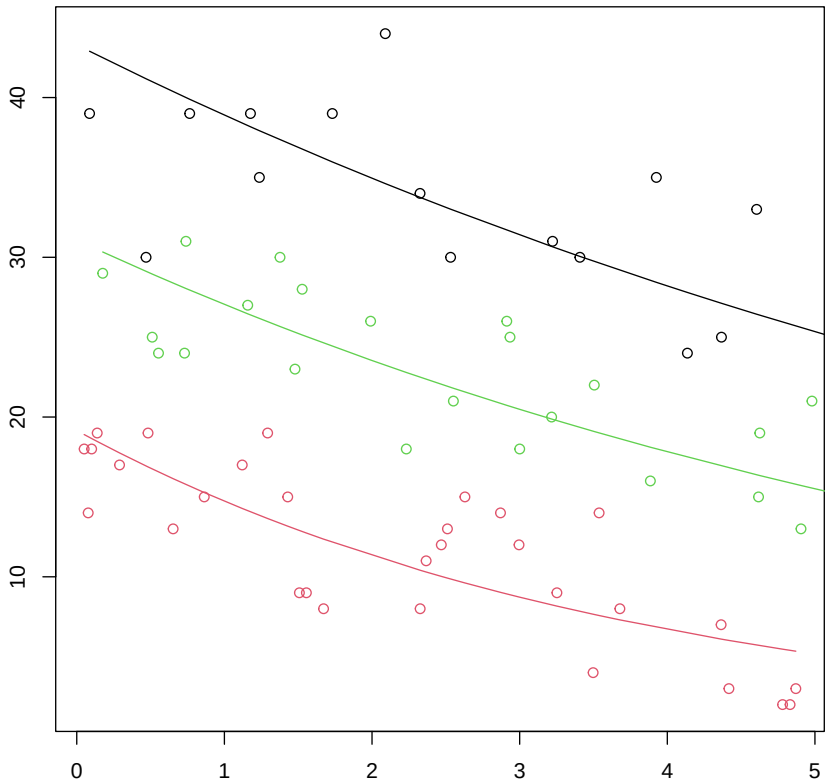
Exemple : nombre d'espèces végétales

Effectifs moyens prédits $\hat{y}_i = \lambda_{\hat{\beta}}(pH_i, Biomass_i)$ (ici toutes les observations sont distinctes) en fonction des effectifs observés y_i.



Exemple : nombre d'espèces végétales

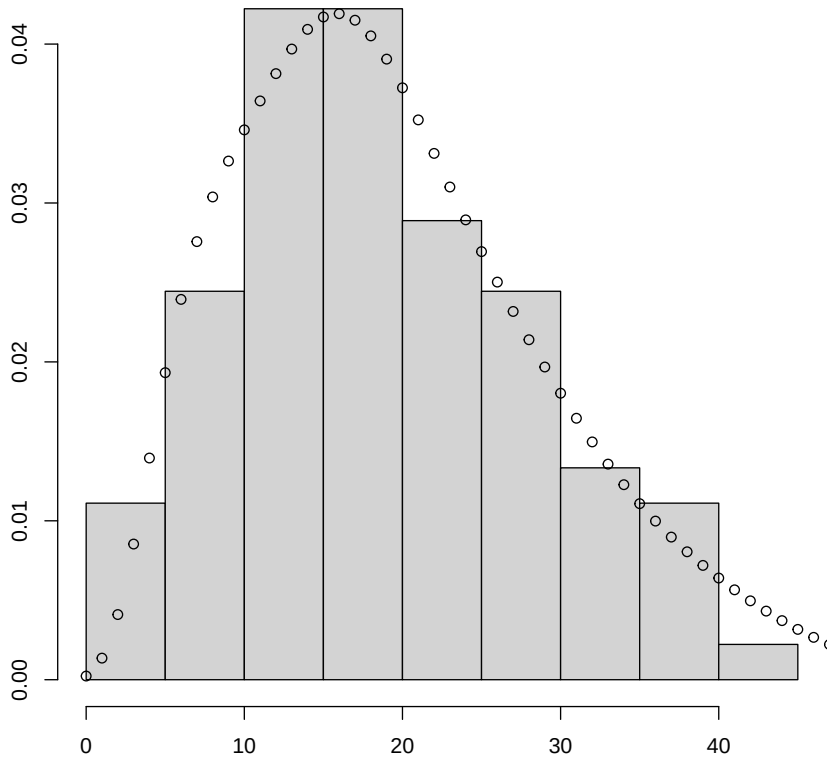
Lignes : effectifs moyens prédits par pH en fonction de la biomasse
Points : effectifs observés par pH en fonction de la biomasse



Noir : pH=basique ; Vert : pH=neutre ; Rouge : pH=acide

Exemple : nombre d'espèces végétales

Histogramme : distribution empirique de Y sur l'échantillon
Points : distribution prédite de Y sachant les x_i observés.



Lorsqu'on modélise $Y|(X=x) \sim \mathcal{P}(\lambda(x))$, on a

$$\mathbb{E}(Y|X=x) = \lambda(x)$$

mais aussi

$$\mathbb{V}(Y|X=x) = \lambda(x).$$

- Cette contrainte est une limite du modèle de Poisson.
- Certaines données sont sur-dispersées, dans le sens où $\mathbb{V}(Y|X=x) > \mathbb{E}(Y|X=x)$.
- Plus rarement, on peut trouver des données sous-dispersées.
- En cas de sur-dispersion, la variance estimée des estimateurs est sous-estimée (inversement pour la sous-dispersion), ce qui fausse l'inférence.

Comment détecter la sur-dispersion ?

- en supposant que $\mathbb{V}(Y|X=x) = \phi\,\mathbb{E}(Y|X=x)$ où $\phi > 0$, on peut estimer ϕ par

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

  et tester si $\phi = 1$ ou non (si $\phi = 1$, $\hat{\phi} \sim \mathcal{N}(1, 1/n)$ lorsque $n \to \infty$).
- on peut ajuster un modèle binomial négatif (cf la suite), et tester s'il est meilleur que le modèle de Poisson.

Solutions possibles en cas de sur-dispersion :

- Ajouter des régresseurs qui "absorbent" la sur-dispersion
- Utiliser une autre loi que la loi de Poisson...

Le modèle quasi-Poisson

Le modèle quasi-Poisson consiste en

1. Estimer $\mathbb{E}(Y|X=x) = \lambda_\beta(x)$ avec $\lambda_\beta(x) = \exp(x'\beta)$ de la même manière qu'avec un modèle de Poisson (même vraisemblance).
2. Estimer ϕ comme dans le slide précédent.
3. Ajuster l'estimation de la variance de $\hat{\beta}$ en tenant compte de $\hat{\phi}$.

Ainsi par rapport à un modèle log-linéaire de Poisson :

- Les coefficients estimés $\hat{\beta}$ sont identiques.
- Les prévisions $\hat{y}_i = \exp(\lambda_{\hat{\beta}}(x_i))$ sont identiques.
- Seules les écarts-types d'estimation diffèrent,
- et donc possiblement la significativité des coefficients.
- La procédure d'estimation ne s'appuie pas sur la "vraie" vraisemblance (à cause de ϕ) : on n'a donc pas accès à L_{mod}.

Exemple : nombre d'espèces végétales

```
glm(Species~ pH+ Biomass+ pH:Biomass, family=quasipoisson)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.76812    0.06144  61.332  < 2e-16 ***
pHlow         -0.81557    0.10268  -7.943 7.90e-12 ***
pHmid         -0.33146    0.09203  -3.602 0.000534 ***
Biomass       -0.10713    0.01247  -8.590 3.97e-13 ***
pHlow:Biomass -0.15503    0.03997  -3.878 0.000208 ***
pHmid:Biomass -0.03189    0.02304  -1.384 0.169985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.9970074)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: NA
```

- $\hat{\phi} = 0.997 \approx 1$ donc il n'y avait pas de souci de sur-dispersion
- La `Residual deviance` est fausse : c'est celle du modèle de Poisson.

Loi binomiale négative

La loi binomiale négative NB peut être une alternative à la loi de Poisson.

Elle dépend de 2 paramètres :

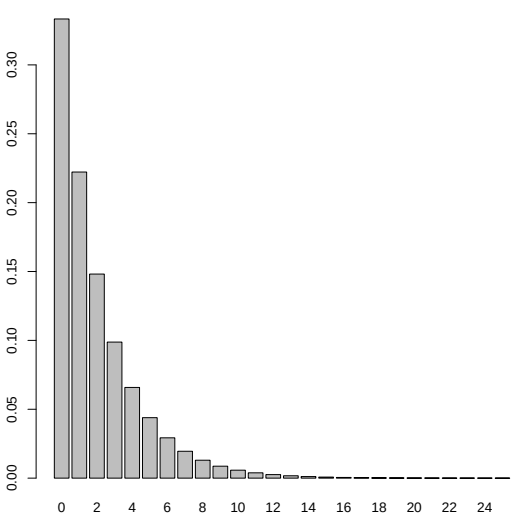
- son espérance $\lambda > 0$
- le "nombre de succès" (ou size) $\theta > 0$

Si $Y \sim NB(\lambda, \theta)$, alors pour tout $k \in \mathbb{N}$,

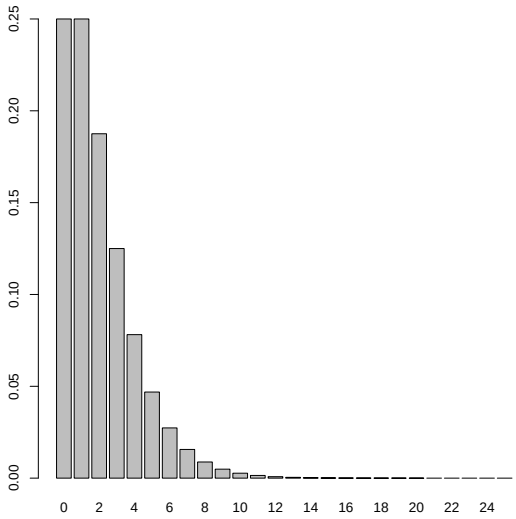
$$\mathbb{P}(Y = k) = \frac{\Gamma(k+\theta)}{\Gamma(k+1)\Gamma(\theta)} \left(\frac{\lambda}{\lambda+\theta}\right)^k \left(\frac{\theta}{\lambda+\theta}\right)^\theta.$$

- L'espérance de $NB(\lambda, \theta)$ vaut λ.
- La variance de $NB(\lambda, \theta)$ vaut $\lambda + \lambda^2/\theta$.
- Cette loi peut donc modéliser la sur-dispersion (mais pas la sous-dispersion).
- Si $\theta \to +\infty$, $NB(\lambda, \theta) \approx \mathcal{P}(\lambda)$.
- La loi de Poisson est donc un cas particulier de la loi NB.

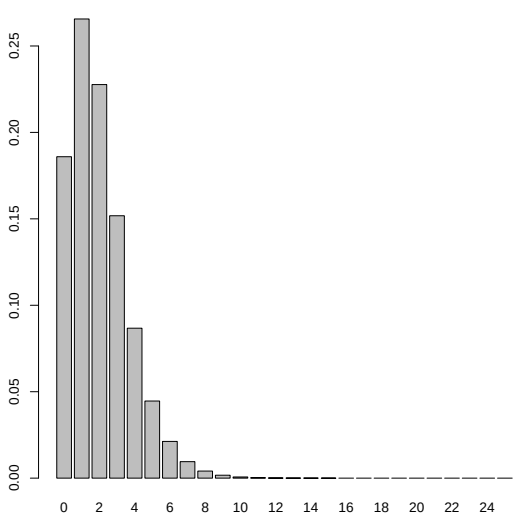
Loi binomiale négative



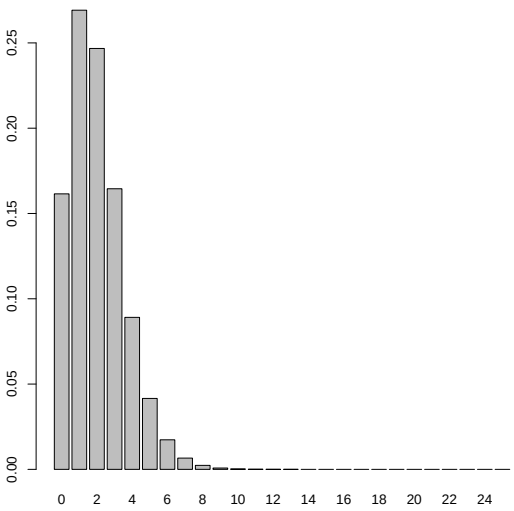
$NB(2,1)$



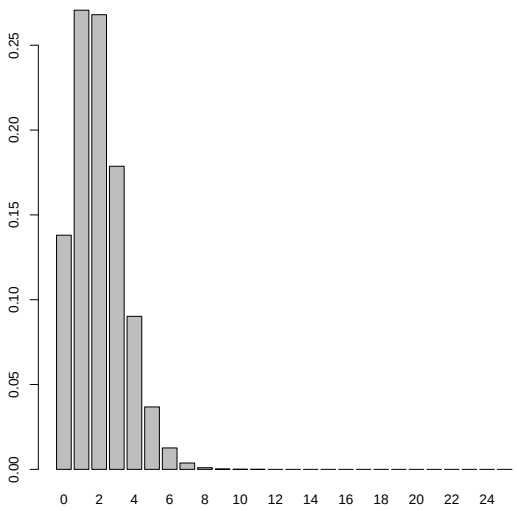
$NB(2,2)$



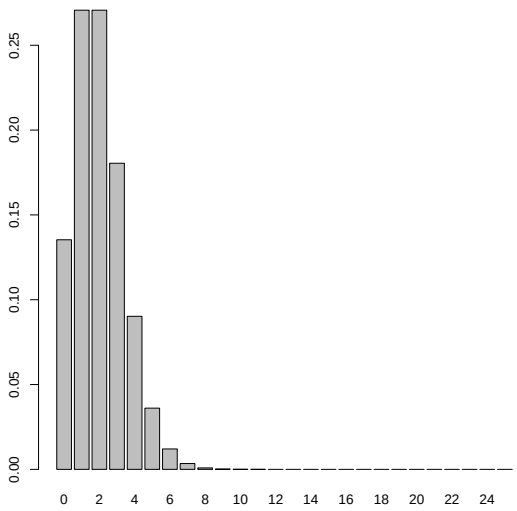
$NB(2,5)$



$NB(2,10)$



$NB(2,100)$



$\mathcal{P}(2)$

Le modèle GLM binomial négatif suppose que

$$Y|(X = x) \sim NB(\lambda(x), \theta), \quad \text{où } \lambda(x) = \exp(x'\beta).$$

L'estimation de β et θ se fait par maximum de vraisemblance

Tous les outils d'inférence usuels sont disponibles :

- Tests de Wald ;
- Deviance ;
- AIC, BIC.

En particulier :

- Si $\hat{\theta}$ est grand, cela revient au modèle de Poisson.
- On peut tester l'intérêt du modèle binomial par rapport au modèle de Poisson en inspectant $\hat{\theta}$, ou en comparant leur critère AIC et BIC.

Exemple : nombre d'espèces végétales

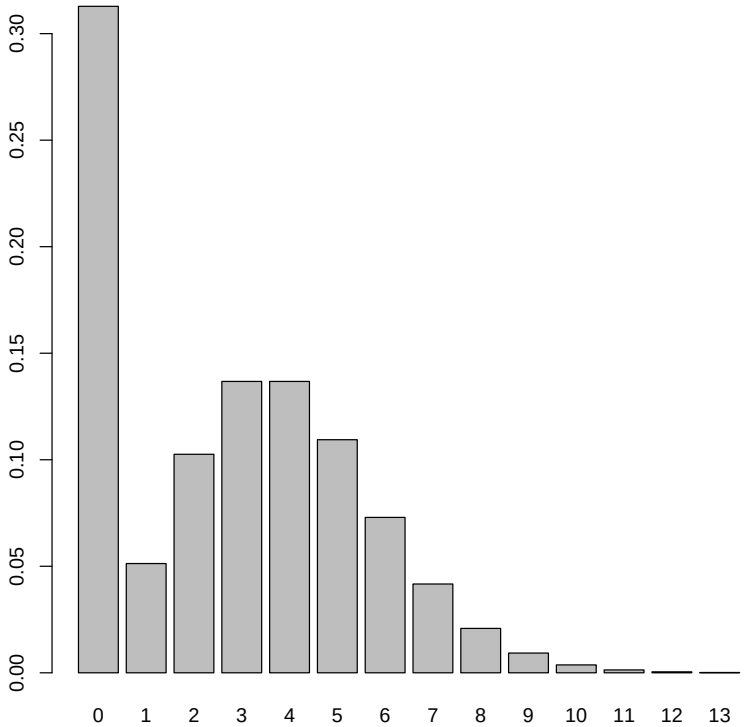
```
                glm.nb(Species~ pH+ Biomass+ pH:Biomass)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76813    0.06154  61.235  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.930 2.19e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow:Biomass -0.15503    0.04003  -3.873 0.000108 ***
pHmid:Biomass -0.03189    0.02308  -1.382 0.166978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(212058.3) family taken to be 1)

    Null deviance: 452.307  on 89  degrees of freedom
Residual deviance:  83.194  on 84  degrees of freedom
AIC: 516.39
```

- $\hat{\theta} = 212058.3$ donc le modèle est équivalent au modèle de Poisson.
- Cela est confirmé via l'AIC.

Lorsque Y est une variable de comptage, il n'est pas rare que $Y = 0$ apparaisse très souvent dans l'échantillon.

Exemples

- Y : quantité de pluie (en mm) tombée chaque jour.
- Y : quantité d'alcool (en verres) consommée chaque semaine.



Un modèle de Poisson ou Binomial Négatif n'est pas adapté à ce type de situations.

Il y a en général deux "populations" qui expliquent ce phénomène :

- celle pour qui $Y = 0$ de façon systématique
- celle pour qui $Y \geq 0$ (ou $Y > 0$)

On peut envisager deux modélisations dans cet esprit :

- Le modèle à inflation de zéros ($Y = 0$ versus $Y \geq 0$)
- Le modèle de Hurdle ($Y = 0$ versus $Y > 0$)

On présente par la suite le **modèle à inflation de zéros**.

Dans le modèle ZIP (Zero-Inflated Poisson), on suppose que

$$\begin{cases} Y|(X=x) = 0 & \text{avec probabilité } q(x), \\ Y|(X=x) \sim \mathcal{P}(\lambda(x)) & \text{avec probabilité } 1-q(x), \end{cases}$$

où

$$\lambda(x) = \exp(x'\beta) \quad \text{et} \quad q(x) = \mathit{logit}^{-1}(x'\gamma).$$

Ainsi deux populations se mélangent :

- Pour l'une Y vaut toujours 0, pour l'autre $Y \in \mathbb{N}$.
- Un modèle logistique explique l'appartenance à l'une ou l'autre population.
- Un modèle log-linéaire de Poisson est utilisé pour la seconde.
- On peut tout à fait imaginer que les régresseurs expliquant $\lambda(x)$ sont différents de ceux expliquant $q(x)$ (il suffit que certains coefficients de β et/ou γ soient nuls)

Avec ce modèle, on a donc

$$\mathbb{P}(Y = 0|X = x) = q_\gamma(x) + (1 - q_\gamma(x))e^{-\lambda_\beta(x)}$$

$$\mathbb{P}(Y = k|X = x) = (1 - q_\gamma(x))e^{-\lambda_\beta(x)} \frac{\lambda_\beta(x)^k}{k!}, \quad k = 1, 2, \ldots$$

où $\lambda_\beta(x) = \exp(x'\beta)$ et $q_\gamma(x) = logit^{-1}(x'\gamma)$.

On en déduit

$$\mathbb{E}(Y|X = x) = (1 - q_\gamma(x))\lambda_\beta(x).$$

Pour l'inférence :

- On peut écrire la log-vraisemblance en fonction de β et γ
- On obtient les estimateurs par maximum de vraisemblance
- Et on a accès aux outils d'inférence habituels

Sous R : fonction `zeroinfl` du package `pscl`.

De même, le modèle ZINB (Zero-Inflated Negative Binomial) est

$$\begin{cases} Y|(X=x) = 0 & \text{avec probabilité } q(x), \\ Y|(X=x) \sim NB(\lambda(x), \theta) & \text{avec probabilité } 1-q(x), \end{cases}$$

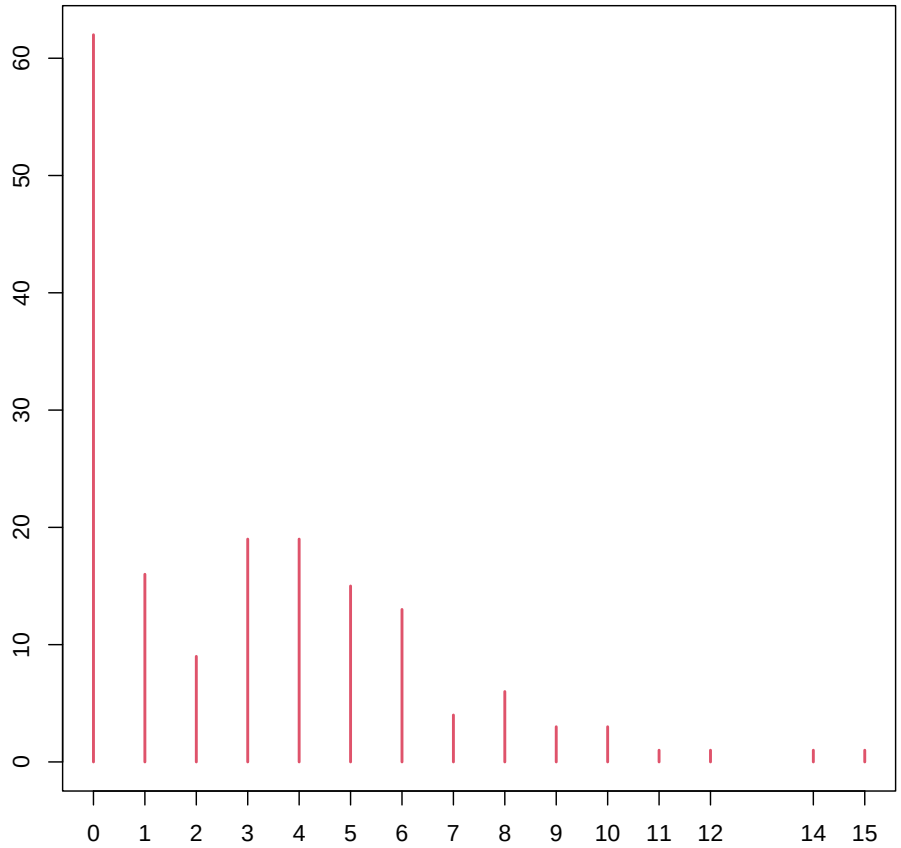
où $\theta > 0$ et

$$\lambda(x) = \exp(x'\beta) \quad \text{et} \quad q(x) = logit^{-1}(x'\gamma).$$

Sous R : fonction `zeroinfl` avec l'option `dist="negbin"`.

Example : Horshoe Crabs

Nombre de satellites mâles sur des limules femelles.



On souhaite modéliser le nombre de satellites (`satell`) en fonction du poids de la limule (`weight`) et de sa couleur (`color`, de 1 à 4, $\approx$ âge).

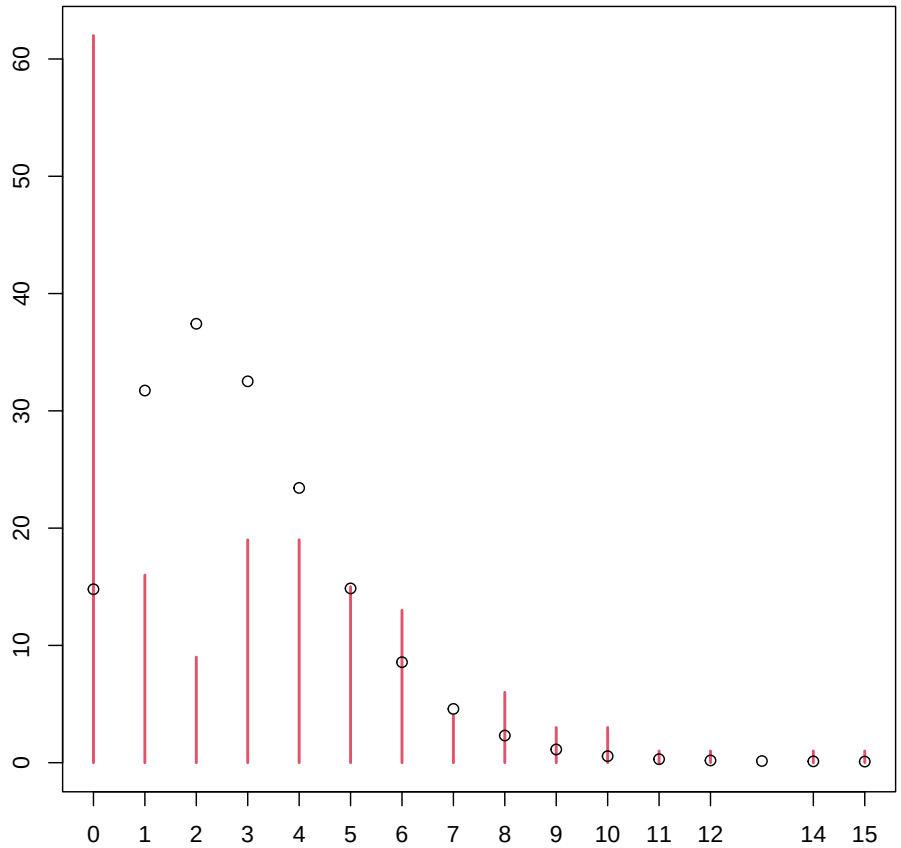
Modèle log-linéaire de Poisson :

```
glm(satell~weight+color,family='poisson',data=crabs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.08855    0.25443   0.348  0.72783
weight       0.54588    0.06749   8.088 6.05e-16 ***
color       -0.17282    0.06155  -2.808  0.00499 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 552.79  on 170  degrees of freedom
AIC: 914.09
```

- Le modèle est significatif mais l'ajustement est médiocre (la déviance est très élevée)

Modèle log-linéaire de Poisson :

En rouge : distribution empirique de `satell` sur l'échantillon
Points : distribution prédite sachant les x_i observés.



- L'ajustement est effectivement mauvais.

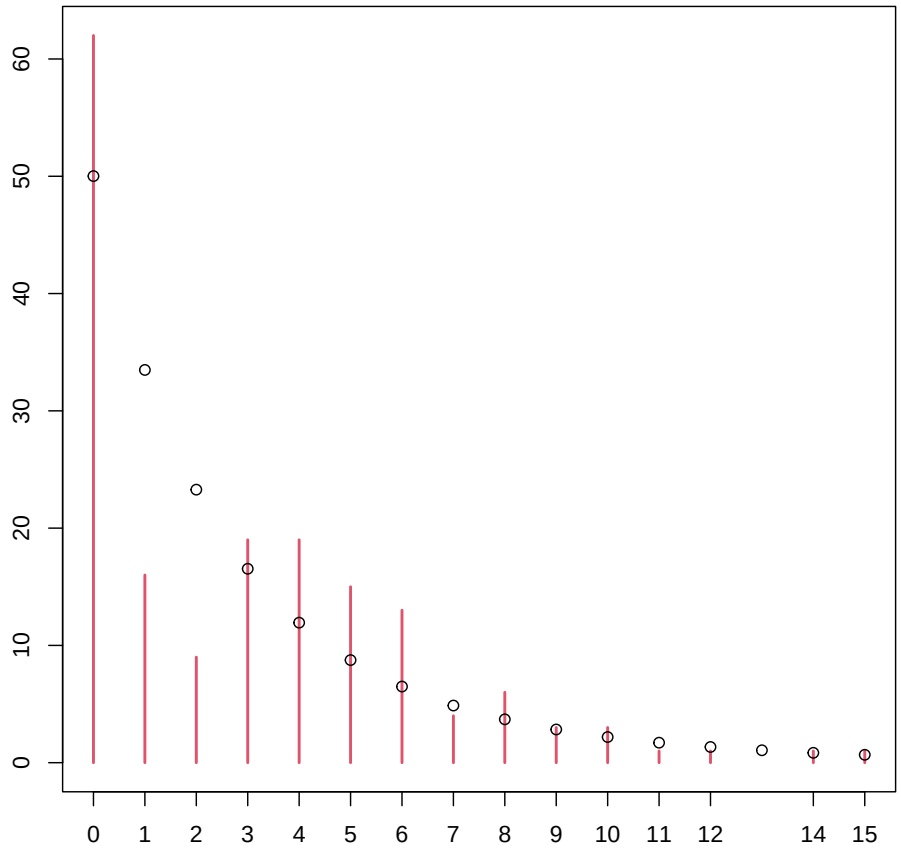
Modèle Binomial Négatif :

```
glm.nb(satell~weight+color,data=crabs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3220     0.5540  -0.581    0.561
weight        0.7072     0.1612   4.387 1.15e-05 ***
color        -0.1734     0.1199  -1.445    0.148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.9555) family taken to be 1)

    Null deviance: 219.50  on 172  degrees of freedom
Residual deviance: 196.64  on 170  degrees of freedom
AIC: 754.45
```

- C'est mieux... (au passage : color ne semble pas significative)

Modèle Binomial Négatif :

En rouge : distribution empirique de `satell` sur l'échantillon
Points : distribution prédite sachant les x_i observés.



- ... mais toujours pas satisfaisant.

Modèle ZIP (Poisson à inflations de zéros) :

```
zeroinfl(satell∼weight | weight+color,dist="poisson",data=crabs)
```

On obtient :

```
Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.00152    0.20793   4.817 1.46e-06 ***
weight       0.19020    0.07572   2.512    0.012 *

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9621     1.1448   1.714   0.0866 .
weight       -1.6630     0.3943  -4.218 2.47e-05 ***
color         0.5329     0.2305   2.312   0.0208 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Number of iterations in BFGS optimization: 9
Log-likelihood: -360.8 on 5 Df
```

Modèle ZIP (Poisson à inflations de zéros) :
Dans la sortie précédente :

- On voit le résultat d'estimation expliquant le mélange des deux "populations" formant le modèle.
- La population qui peut prendre des valeurs $Y \in \mathbb{N}$ suit un modèle log-linéaire de Poisson d'espérance λ avec

$$\lambda(weight) = \exp(1 + 0.19 weight)$$

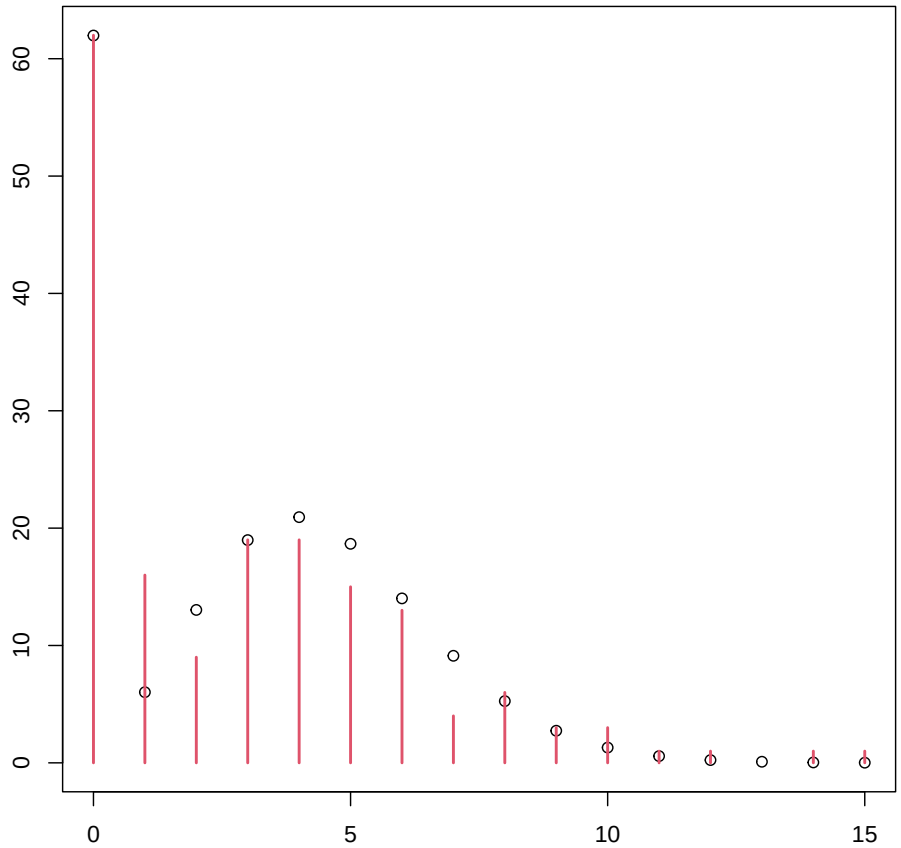
- La probabilité q d'appartenir à l'autre population (pour laquelle Y vaut toujours 0) vaut

$$q(weight, color) = logit^{-1}(1.96 - 1.66 weight + 0.53 color)$$

- Et donc la probabilité d'appartenir à la population pour qui $Y \in \mathbb{N}$ est $1 - q(weight, color)$.

Modèle ZIP (Poisson à inflations de zéros) :

En rouge : distribution empirique de `satell` sur l'échantillon
Points : distribution prédite sachant les x_i observés.



- Le résultat est plus convaincant.

Modèle ZINB (NB à inflations de zéros) :

```
zeroinfl(satell∼weight | weight+color,dist="negbin",data=crabs)
```

On obtient :

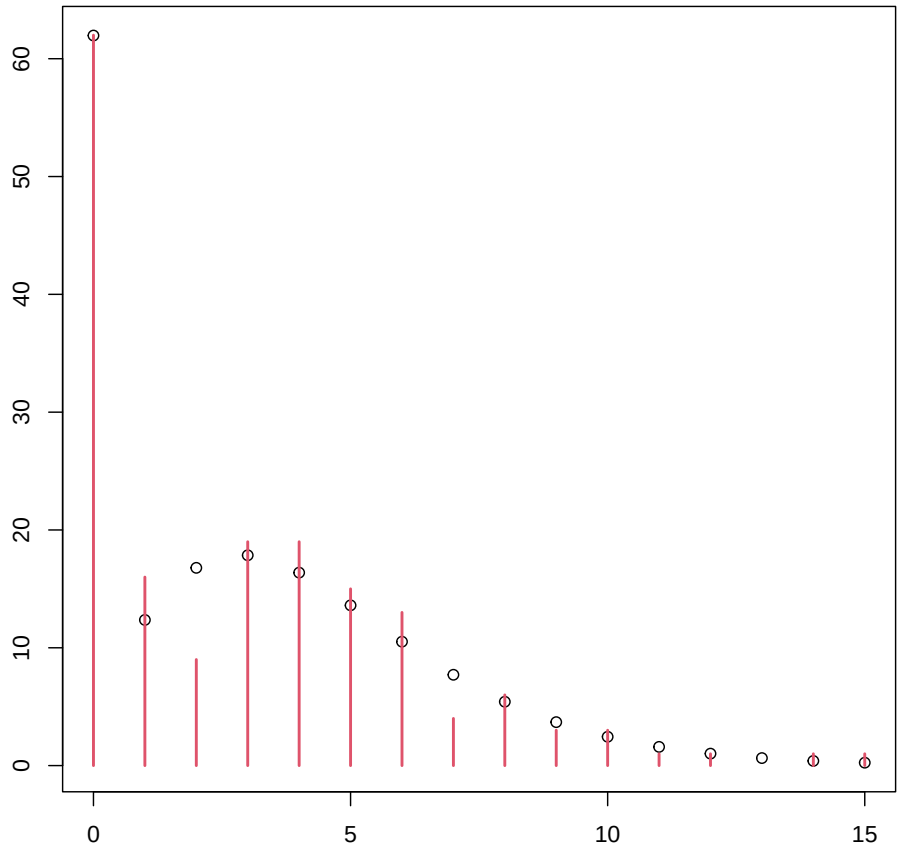
```
Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8961     0.3070   2.919  0.00351 **
weight        0.2169     0.1125   1.928  0.05383 .
Log(theta)    1.5802     0.3574   4.422 9.79e-06 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8663     1.2415   1.503    0.133
weight       -1.7531     0.4429  -3.958 7.55e-05 ***
color         0.5985     0.2572   2.326    0.020 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 4.8558
Number of iterations in BFGS optimization: 11
Log-likelihood: -349.9 on 6 Df
```

Modèle ZINB (NB à inflations de zéros) :

- La sortie se lit de la même manière que pour le modèle ZIP
- Il y a le paramètre θ en plus, issu de la loi NB
- Ce dernier n'est pas "infini" : le modèle ne se réduit donc pas au modèle ZIP.
- Un test de déviance montre que ZINB est préférable à ZIP
- Cela est confirmé également par les critères AIC et BIC.

Modèle ZINB (NB à inflations de zéros) :

En rouge : distribution empirique de `satell` sur l'échantillon
Points : distribution prédite sachant les x_i observés.



- L'ajustement est satisfaisant

Conclusion de l'étude : d'après le modèle ZINB,

- Une partie des limules (les plus petites et les plus âgées) n'ont pas de satellite mâle.
- La probabilité d'appartenir à cette population est estimée à

$$q(weight, color) = logit^{-1}(1.87 - 1.75 weight + 0.60 color)$$

- Pour l'autre partie des limules, elles ont en moyenne d'autant plus de satellites qu'elles sont grosses. Cette moyenne est estimée à

$$\lambda(weight) = \exp(0.90 + 0.22 weight).$$

- La distribution précise du nombre de satellites pour cette population peut être modélisée par une $NB(\lambda(weight), \theta)$ où $\theta = 4.86$.

C'est la fin du cours.

Merci de votre attention !

Références

- "Régression avec R", P-A. Cornillon, E. Matzner-Løber
  $\longrightarrow$ *Livre en français, très accessible, en lien avec les 3 premiers chapitres*

- "Le modèle linéaire par l'exemple", J.-M. Azais, J.-M. Bardet.
  $\longrightarrow$ *Livre en français, en lien avec les 3 premiers chapitres : des discussions intéressantes sur l'enjeu des hypothèses, et des résultats théoriques fins.*

- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
  $\longrightarrow$ *Grand classique sur les méthodes de machine learning, y compris les méthodes vues dans ce cours. Exemples avec R.*

- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.
  $\longrightarrow$ *Grand classique également. Version plus théorique (et plus complète) que le précédent.*

Références (suite)

- Agresti, A. Foundations of Linear and Generalized Linear Models, Wiley.
  $\longrightarrow$ *Livre classique sur le sujet, en lien avec le chapitre 4*

- Antoniadis, A. Berruyer J. et Carmona R. Régression non linéaire et applications, Economica.
  $\longrightarrow$ *Résultats théoriques complets, en lien avec le chapitre 4*

- Dobson, A.J., Barnett, A.G. An Introduction to Generalized Linear Models, CRC Press.
  $\longrightarrow$ *Des exemples en R, en lien avec le chapitre 4*

- Hosmer, D. et Lemeshow S. Applied Logistic Regression, Wiley.
  $\longrightarrow$ *La régression logistique en applications, en long et en large*