Régression logistique et scoring

L. Rouvière
laurent.rouviere@univ-rennes2.fr

Janvier 2017

Plan du cours

Première partie I

Introduction aux GLM

Modèle

Qu'est-ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$ avec

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Modèle

Qu'est-ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$ avec

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Modèle

Qu'est-ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$ avec

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Modèle

Qu'est-ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$ avec

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Modèle

Qu'est-ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$ avec

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Exemple 1

- On souhaite tester l'efficacité d'un nouveau traitement à l'aide d'un essai clinique.
- On traite $n = 100$ patients atteints de la pathologie.
- A l'issue de l'étude, 72 patients sont guéris.

- Soit p_0 la probabilité de guérison suite au traitement en question.
- On est tentés de conclure $p_0 \approx 0.72$.

Un tel résultat n'a cependant guère d'intêret si on n'est pas capable de préciser l'erreur susceptible d'être commise par cette estimation.

Exemple 1

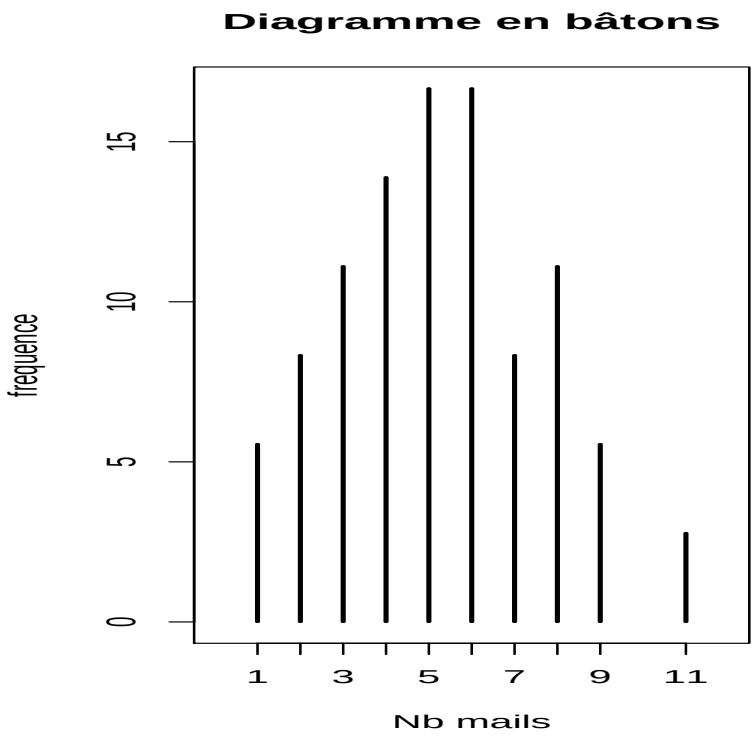
- On souhaite tester l'efficacité d'un nouveau traitement à l'aide d'un essai clinique.
- On traite $n = 100$ patients atteints de la pathologie.
- A l'issue de l'étude, 72 patients sont guéris.

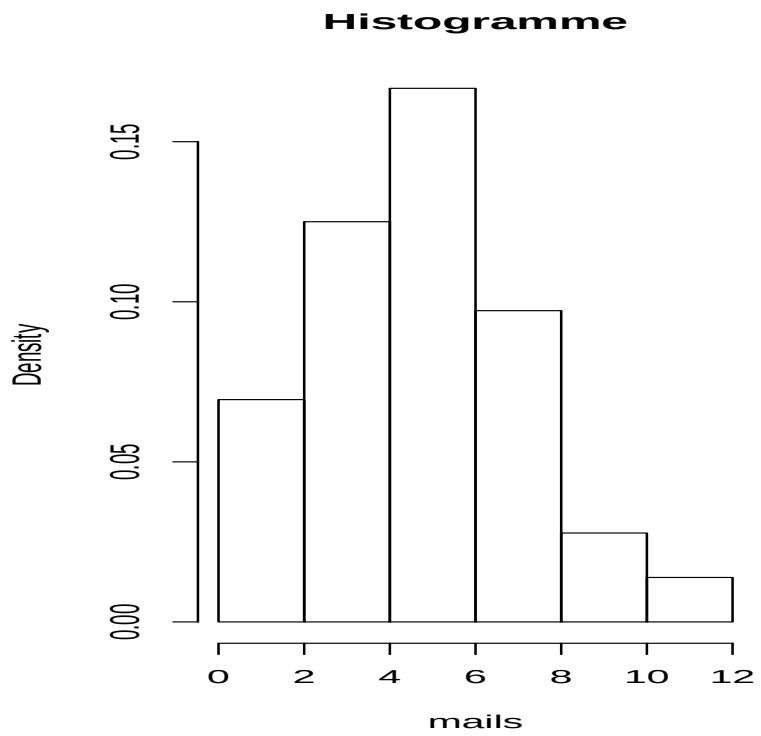
- Soit p_0 la probabilité de guérison suite au traitement en question.
- On est tentés de conclure $p_0 \approx 0.72$.

Un tel résultat n'a cependant guère d'intêret si on n'est pas capable de préciser l'erreur susceptible d'être commise par cette estimation.

Exemple 2

- On s'intéresse au nombre de mails reçus par jour par un utilisateur pendant 36 journées.
- $\bar{x} = 5.22$, $S_n^2 = 5.72$.

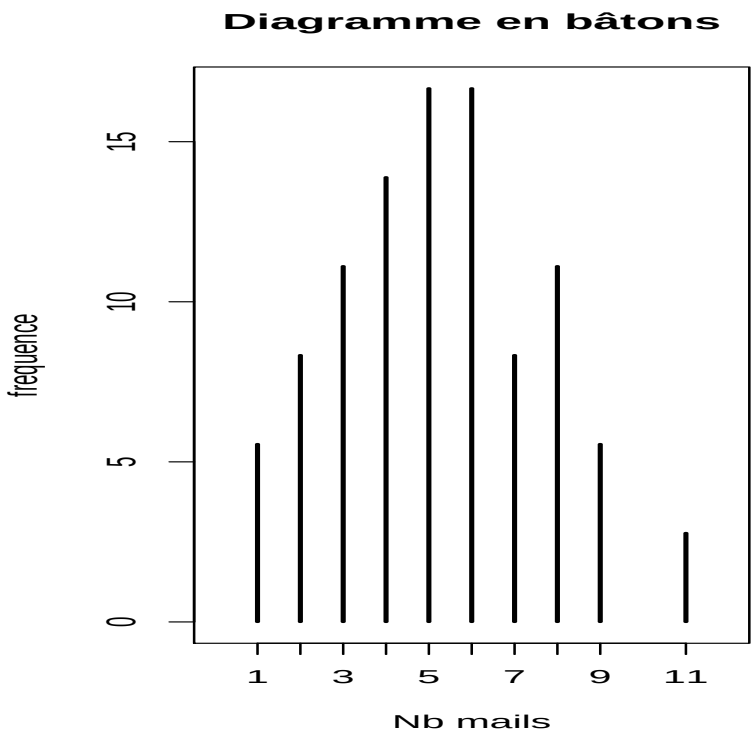


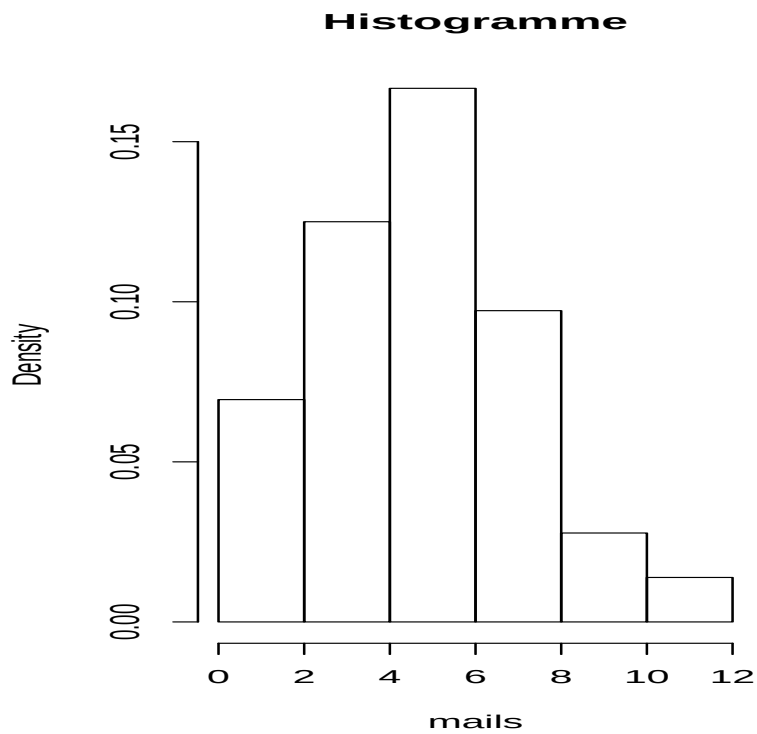


Quelle est la probabilité de recevoir plus de 5 mails dans une journée ?

Exemple 2

- On s'intéresse au nombre de mails reçus par jour par un utilisateur pendant 36 journées.
- $\bar{x} = 5.22$, $S_n^2 = 5.72$.

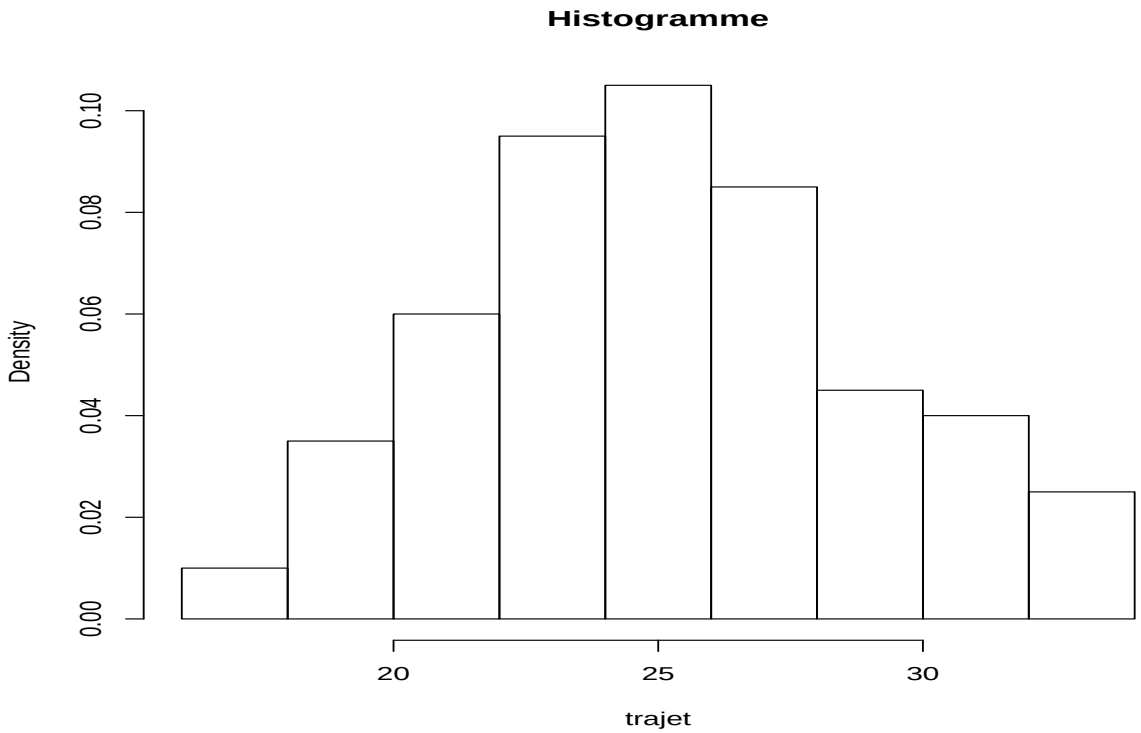




Quelle est la probabilité de recevoir plus de 5 mails dans une journée ?

Exemple 3

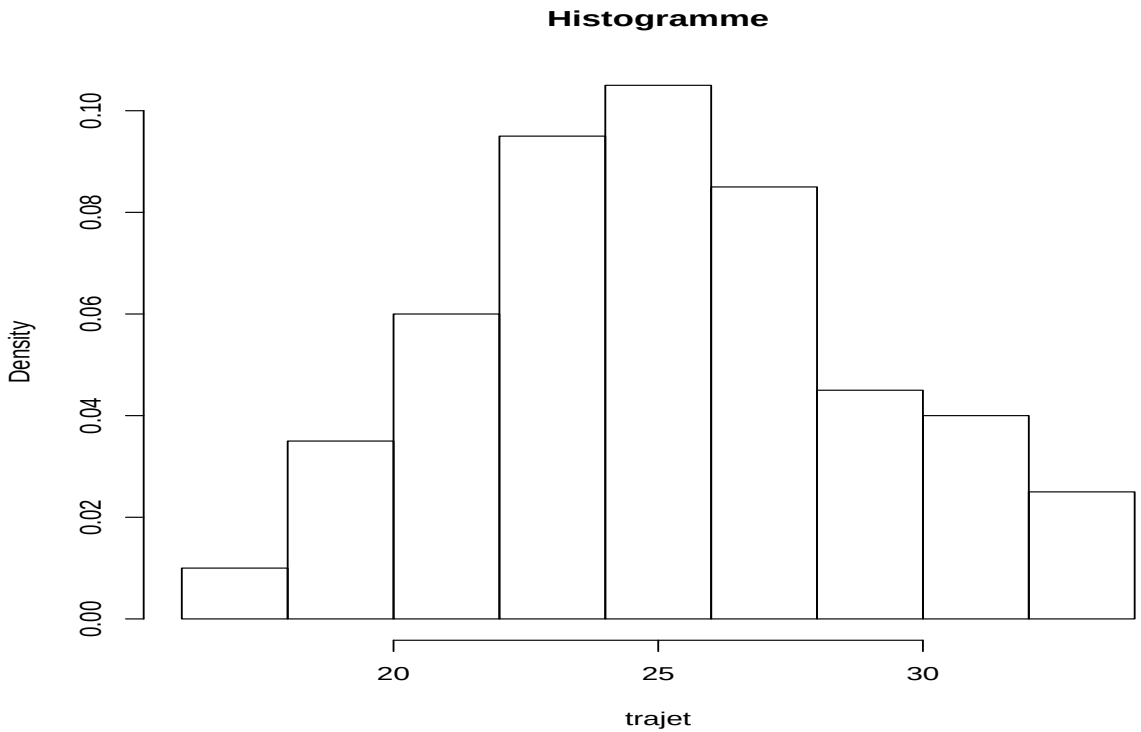
- Durée de trajet domicile-travail.
- On dispose de $n = 100$ mesures : $\bar{x} = 25.1$, $S_n^2 = 14.46$.



J'ai une réunion à 8h30, quelle est la probabilité que j'arrive en retard si je pars de chez moi à 7h55 ?

Exemple 3

- Durée de trajet domicile-travail.
- On dispose de $n = 100$ mesures : $\bar{x} = 25.1$, $S_n^2 = 14.46$.



J'ai une réunion à 8h30, quelle est la probabilité que j'arrive en retard si je pars de chez moi à 7h55 ?

Problème

- Nécessité de se dégager des observations $x_1, \ldots, x_n$ pour répondre à de telles questions.
- Si on mesure la durée du trajet pendant 100 nouveaux jours, on peut en effet penser que les nouvelles observations ne seront pas exactement les mêmes que les anciennes.

Idée

Considérer que les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$.

Attention

X_i est une variable aléatoire et x_i est une réalisation de cette variable, c'est-à-dire un nombre !

Problème

- Nécessité de se dégager des observations $x_1, \ldots, x_n$ pour répondre à de telles questions.
- Si on mesure la durée du trajet pendant 100 nouveaux jours, on peut en effet penser que les nouvelles observations ne seront pas exactement les mêmes que les anciennes.

Idée

Considérer que les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$.

Attention

X_i est une variable aléatoire et x_i est une réalisation de cette variable, c'est-à-dire un nombre !

Problème

- Nécessité de se dégager des observations $x_1, \ldots, x_n$ pour répondre à de telles questions.
- Si on mesure la durée du trajet pendant 100 nouveaux jours, on peut en effet penser que les nouvelles observations ne seront pas exactement les mêmes que les anciennes.

Idée

Considérer que les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$.

Attention

X_i est une variable aléatoire et x_i est une réalisation de cette variable, c'est-à-dire un nombre !

Variables aléatoires

Définition

Une variable aléatoire réelle est une application

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

telle que

$$\forall B \in \mathcal{B}(\mathbb{R}), X^{-1}(B) \in \mathcal{A}.$$

- Lors de la modélisation statistique, l'espace Ω n'est généralement jamais caractérisé.
- Il contient tous les "phénoménes" pouvant expliquer les sources d'aléa (qui ne sont pas explicables...).
- En pratique, l'espace d'arrivée est généralement suffisant.

Variables aléatoires

Définition

Une variable aléatoire réelle est une application

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

telle que

$$\forall B \in \mathcal{B}(\mathbb{R}), X^{-1}(B) \in \mathcal{A}.$$

- Lors de la modélisation statistique, l'espace Ω n'est généralement jamais caractérisé.
- Il contient tous les "phénoménes" pouvant expliquer les sources d'aléa (qui ne sont pas explicables...).
- En pratique, l'espace d'arrivée est généralement suffisant.

Variables aléatoires

Définition

Une variable aléatoire réelle est une application

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

telle que

$$\forall B \in \mathcal{B}(\mathbb{R}), X^{-1}(B) \in \mathcal{A}.$$

- Lors de la modélisation statistique, l'espace Ω n'est généralement jamais caractérisé.
- Il contient tous les "phénoménes" pouvant expliquer les sources d'aléa (qui ne sont pas explicables...).
- En pratique, l'espace d'arrivée est généralement suffisant.

Loi de probabilité

Loi de probabilité

Etant donnée $\mathbf{P}$ une probabilité sur $(\Omega, \mathcal{A})$ et X une variable aléatoire réelle définie sur Ω, on appelle loi de probabilité de X la mesure $\mathbf{P}_X$ définie par

$$\mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) = \mathbf{P}(X \in B) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in B\}) \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Une loi de probabilité est caractérisée par

- sa fonction de répartition : $F_X(x) = \mathbf{P}(X \leq x)$.
- sa densité : $f_x : \mathbb{R} \to \mathbb{R}^+$ telle que $\forall B \in \mathcal{B}(\mathbb{R})$

$$\mathbf{P}_X(B) = \int_B f_X(x)\, dx.$$

Loi de probabilité

Etant donnée $\mathbf{P}$ une probabilité sur $(\Omega, \mathcal{A})$ et X une variable aléatoire réelle définie sur Ω, on appelle loi de probabilité de X la mesure $\mathbf{P}_X$ définie par

$$\mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) = \mathbf{P}(X \in B) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in B\}) \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Une loi de probabilité est caractérisée par

- sa fonction de répartition : $F_X(x) = \mathbf{P}(X \leq x)$.
- sa densité : $f_x : \mathbb{R} \to \mathbb{R}^+$ telle que $\forall B \in \mathcal{B}(\mathbb{R})$

$$\mathbf{P}_X(B) = \int_B f_X(x)\, dx.$$

Un modèle pour l'exemple 1

- On note $x_i = 1$ si le $i^{\text{ème}}$ patient a guéri, 0 sinon.
- On peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de bernoulli de paramètre p_0.
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de guérir (ce qui peut revenir à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi (i.i.d.).

On dit que $X_1, \ldots, X_n$ est un *n*-échantillon de variables aléatoires indépendantes de même loi $B(p_0)$.

Un modèle pour l'exemple 1

- On note $x_i = 1$ si le $i^{\text{ème}}$ patient a guéri, 0 sinon.
- On peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de bernoulli de paramètre p_0.
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de guérir (ce qui peut revenir à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi (i.i.d.).

On dit que $X_1, \ldots, X_n$ est un n-échantillon de variables aléatoires indépendantes de même loi $B(p_0)$.

Un modèle pour l'exemple 1

- On note $x_i = 1$ si le $i^{\text{ème}}$ patient a guéri, 0 sinon.
- On peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de bernoulli de paramètre p_0.
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de guérir (ce qui peut revenir à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi (i.i.d.).

On dit que $X_1, \ldots, X_n$ est un *n*-échantillon de variables aléatoires indépendantes de même loi $B(p_0)$.

Définitions

Modèle

On appelle **modèle statistique** tout triplet $(\mathcal{H}, \mathcal{A}, \mathcal{P})$ où

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

Le problème du statisticien

- n variables aléatoires i.i.d. $X_1, \ldots, X_n$ de loi $\mathbf{P}$.
- Trouver une famille de lois $\mathcal{P}$ susceptible de contenir $\mathbf{P}$.
- Trouver dans $\mathcal{P}$ une loi qui soit la plus proche de $\mathbf{P}$

Définitions

Modèle

On appelle **modèle statistique** tout triplet $(\mathcal{H}, \mathcal{A}, \mathcal{P})$ où

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- $\mathcal{A}$ est une tribu sur $\mathcal{H}$;
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

Le problème du statisticien

- n variables aléatoires i.i.d. $X_1, \ldots, X_n$ de loi $\mathbf{P}$.
- Trouver une famille de lois $\mathcal{P}$ susceptible de contenir $\mathbf{P}$.
- Trouver dans $\mathcal{P}$ une loi qui soit la plus proche de $\mathbf{P}$

Exemples

	$\mathcal{H}$	$\mathcal{A}$	$\mathcal{P}$
Exemple 1	$\{0,1\}$	$\mathcal{P}(\{0,1\})$	$\{B(p), p \in [0,1]\}$
Exemple 2	$\mathbb{N}$	$\mathcal{P}(\mathbb{N})$	$\{\mathcal{P}(\lambda), \lambda > 0\}$
Exemple 3	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$	$\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$

Paramétrique versus non paramétrique

Définition

- Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.
- Si $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$ où $\mathcal{F}$ est de dimension infinie, on parle de modèle non paramétrique.

Exemple : modèle de densité

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ est un modèle paramétrique.
- $\mathcal{P} = \{\text{densités } f \text{ 2 fois dérivables}\}$ est un modèle non paramétrique.

Le problème sera d'estimer (μ, σ^2) ou f à partir de l'échantillon $X_1, \ldots, X_n$.

Paramétrique versus non paramétrique

Définition

- Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.
- Si $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$ où $\mathcal{F}$ est de dimension infinie, on parle de modèle non paramétrique.

Exemple : modèle de densité

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ est un modèle paramétrique.
- $\mathcal{P} = \{\text{densités } f \text{ 2 fois dérivables}\}$ est un modèle non paramétrique.

Le problème sera d'estimer (μ, σ^2) ou f à partir de l'échantillon $X_1, \ldots, X_n$.

Paramétrique versus non paramétrique

Définition

- Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.
- Si $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$ où $\mathcal{F}$ est de dimension infinie, on parle de modèle non paramétrique.

Exemple : modèle de densité

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ est un modèle paramétrique.
- $\mathcal{P} = \{\text{densités } f \text{ 2 fois dérivables}\}$ est un modèle non paramétrique.

Le problème sera d'estimer (μ, σ^2) ou f à partir de l'échantillon $X_1, \ldots, X_n$.

Paramétrique versus non paramétrique

Définition

- Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.
- Si $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$ où $\mathcal{F}$ est de dimension infinie, on parle de modèle non paramétrique.

Exemple : modèle de densité

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ est un modèle paramétrique.
- $\mathcal{P} = \{\text{densités } f \text{ 2 fois dérivables}\}$ est un modèle non paramétrique.

Le problème sera d'estimer (μ, σ^2) ou f à partir de l'échantillon $X_1, \ldots, X_n$.

Modèle de régression

- On cherche à expliquer une variable Y par p variables explicatives $\mathbf{X}_1, \ldots, \mathbf{X}_p$. On dispose d'un n échantillon i.i.d. $(X_i, Y_i), i = 1, \ldots, n$.

Modèle linéaire (paramétrique)

- On pose

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- Le problème est d'estimer $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Un modèle non paramétrique

- On pose

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$$

où $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.

- Le problème est d'estimer m à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Modèle de régression

- On cherche à expliquer une variable Y par p variables explicatives $\mathbf{X}_1, \ldots, \mathbf{X}_p$. On dispose d'un n échantillon i.i.d. $(X_i, Y_i), i = 1, \ldots, n$.

Modèle linéaire (paramétrique)

- On pose

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- Le problème est d'estimer $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Un modèle non paramétrique

- On pose

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$$

où $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.

- Le problème est d'estimer m à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Modèle de régression

- On cherche à expliquer une variable Y par p variables explicatives $\mathbf{X}_1, \ldots, \mathbf{X}_p$. On dispose d'un n échantillon i.i.d. $(X_i, Y_i), i = 1, \ldots, n$.

Modèle linéaire (paramétrique)

- On pose

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- Le problème est d'estimer $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Un modèle non paramétrique

- On pose

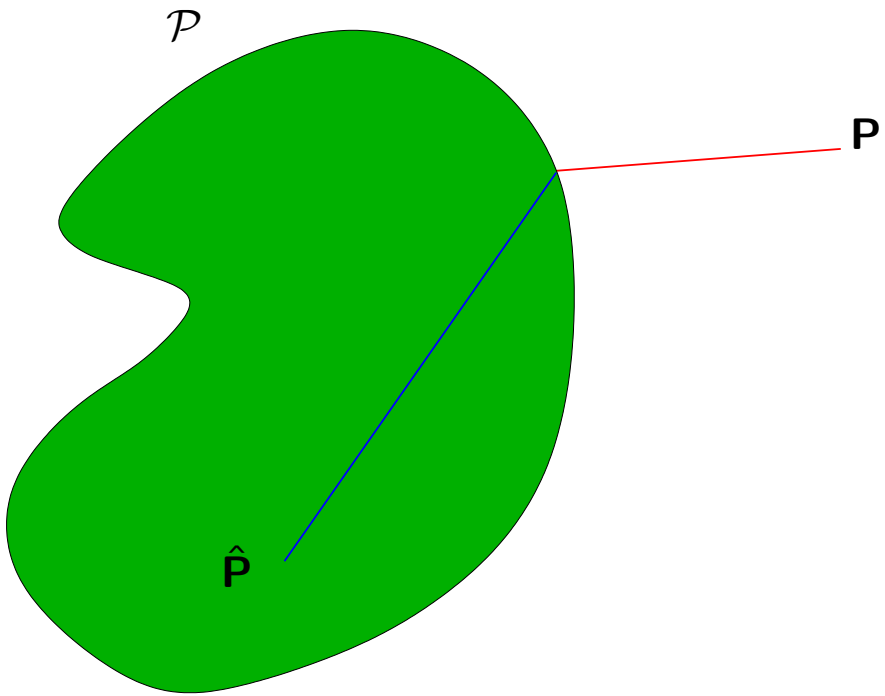
$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$$

où $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.

- Le problème est d'estimer m à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2 types d'erreur

- Poser un modèle revient à choisir une famille de loi candidates pour reconstruire la loi des données $\mathbf{P}$.

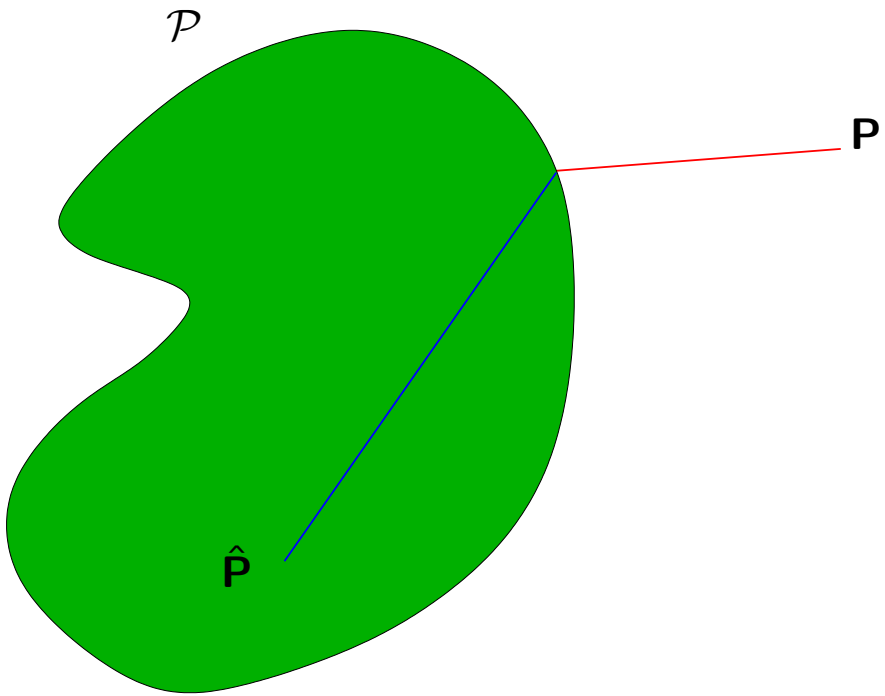


On distingue deux types d'erreurs :

- Erreur d'estimation : erreur commise par le choix d'une loi dans $\mathcal{P}$ par rapport au meilleur choix.
- Erreur d'approximation : erreur commise par le choix de $\mathcal{P}$.

2 types d'erreur

- Poser un modèle revient à choisir une famille de loi candidates pour reconstruire la loi des données $\mathbf{P}$.



On distingue deux types d'erreurs :

- Erreur d'estimation : erreur commise par le choix d'une loi dans $\mathcal{P}$ par rapport au meilleur choix.
- Erreur d'approximation : erreur commise par le choix de $\mathcal{P}$.

La modélisation

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ qui sont le résultats de n expériences aléatoires indépendantes.
2. **Modélisation** : on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi $\mathbf{P}_{\theta_0}$.
3. **Estimation** : chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de $\mathbf{P}_{\theta_0} \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ_0.
4. **"Validation" de modèle** : on revient en arrière et on tente de vérifier si l'hypothèse de l'étape 2 est raisonnable (test d'adéquation, etc...)

La modélisation

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ qui sont le résultats de n expériences aléatoires indépendantes.

2. **Modélisation :** on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi $\mathbf{P}_{\theta_0}$.

3. **Estimation :** chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de $\mathbf{P}_{\theta_0} \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ_0.

4. **"Validation" de modèle :** on revient en arrière et on tente de vérifier si l'hypothèse de l'étape 2 est raisonnable (test d'adéquation, etc...)

La modélisation

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ qui sont le résultats de n expériences aléatoires indépendantes.

2. **Modélisation** : on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi $\mathbf{P}_{\theta_0}$.

3. **Estimation** : chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de $\mathbf{P}_{\theta_0} \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ_0.

4. **"Validation" de modèle** : on revient en arrière et on tente de vérifier si l'hypothèse de l'étape 2 est raisonnable (test d'adéquation, etc...)

La modélisation

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ qui sont le résultats de n expériences aléatoires indépendantes.

2. **Modélisation :** on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi $\mathbf{P}_{\theta_0}$.

3. **Estimation :** chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de $\mathbf{P}_{\theta_0} \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ_0.

4. **"Validation" de modèle** : on revient en arrière et on tente de vérifier si l'hypothèse de l'étape 2 est raisonnable (test d'adéquation, etc...)

Le problème de régression

- On cherche à expliquer une variable Y par p variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.
- Il s'agit de trouver une fonction $m : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx m(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
- Sauf cas (très) particulier, le lien n'est jamais "parfait"

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon.$$

Modèle de régression

- Poser un modèle de régression revient à supposer que la fonction m appartient à un certain espace $\mathcal{M}$.
- Le problème du statisticien sera alors de trouver la "meilleure" fonction dans $\mathcal{M}$ à l'aide d'un n-échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Le problème de régression

- On cherche à expliquer une variable Y par p variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.
- Il s'agit de trouver une fonction $m : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx m(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
- Sauf cas (très) particulier, le lien n'est jamais "parfait"

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon.$$

Modèle de régression

- Poser un modèle de régression revient à supposer que la fonction m appartient à un certain espace $\mathcal{M}$.
- Le problème du statisticien sera alors de trouver la "meilleure" fonction dans $\mathcal{M}$ à l'aide d'un n-échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Le problème de régression

- On cherche à expliquer une variable Y par p variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.
- Il s'agit de trouver une fonction $m : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx m(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
- Sauf cas (très) particulier, le lien n'est jamais "parfait"

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon.$$

Modèle de régression

- Poser un modèle de régression revient à supposer que la fonction m appartient à un certain espace $\mathcal{M}$.
- Le problème du statisticien sera alors de trouver la "meilleure" fonction dans $\mathcal{M}$ à l'aide d'un n-échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Le problème de régression

- On cherche à expliquer une variable Y par p variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.
- Il s'agit de trouver une fonction $m : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx m(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
- Sauf cas (très) particulier, le lien n'est jamais "parfait"

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon.$$

Modèle de régression

- Poser un modèle de régression revient à supposer que la fonction m appartient à un certain espace $\mathcal{M}$.
- Le problème du statisticien sera alors de trouver la "meilleure" fonction dans $\mathcal{M}$ à l'aide d'un n-échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Exemple de modèles

Modèle non paramétrique

- L'espace $\mathcal{M}$ est de dimension infinie.
- **Exemple** : On pose $Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$ où m appartient à l'espace des fonctions continues.

Modèle paramétrique

- L'espace $\mathcal{M}$ est de dimension finie.
- **Exemple** : on suppose que la fonction m est linéaire

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon.$$

Le problème est alors d'estimer $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

- C'est le modèle de **régression linéaire**.

Exemple de modèles

Modèle non paramétrique

- L'espace $\mathcal{M}$ est de dimension infinie.
- **Exemple** : On pose $Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$ où m appartient à l'espace des fonctions continues.

Modèle paramétrique

- L'espace $\mathcal{M}$ est de dimension finie.
- **Exemple** : on suppose que la fonction m est linéaire

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon.$$

Le problème est alors d'estimer $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

- C'est le modèle de **régression linéaire**.

Exemple

- On cherche à **expliquer** ou à **prédire** la concentration en ozone.
- On dispose de $n = 112$ observations de la concentration en ozone ainsi que de 12 autres variables susceptibles d'expliquer cette concentration :
  - Température relevée à différents moments de la journée.
  - Indice de nébulosité relevé à différents moments de la journée.
  - Direction du vent.
  - Pluie.

Question

Comment expliquer (modéliser) la concentration en ozone à l'aide de toutes ces variables ?

Exemple

- On cherche à **expliquer** ou à **prédire** la concentration en ozone.
- On dispose de $n = 112$ observations de la concentration en ozone ainsi que de 12 autres variables susceptibles d'expliquer cette concentration :
  - Température relevée à différents moments de la journée.
  - Indice de nébulosité relevé à différents moments de la journée.
  - Direction du vent.
  - Pluie.

Question

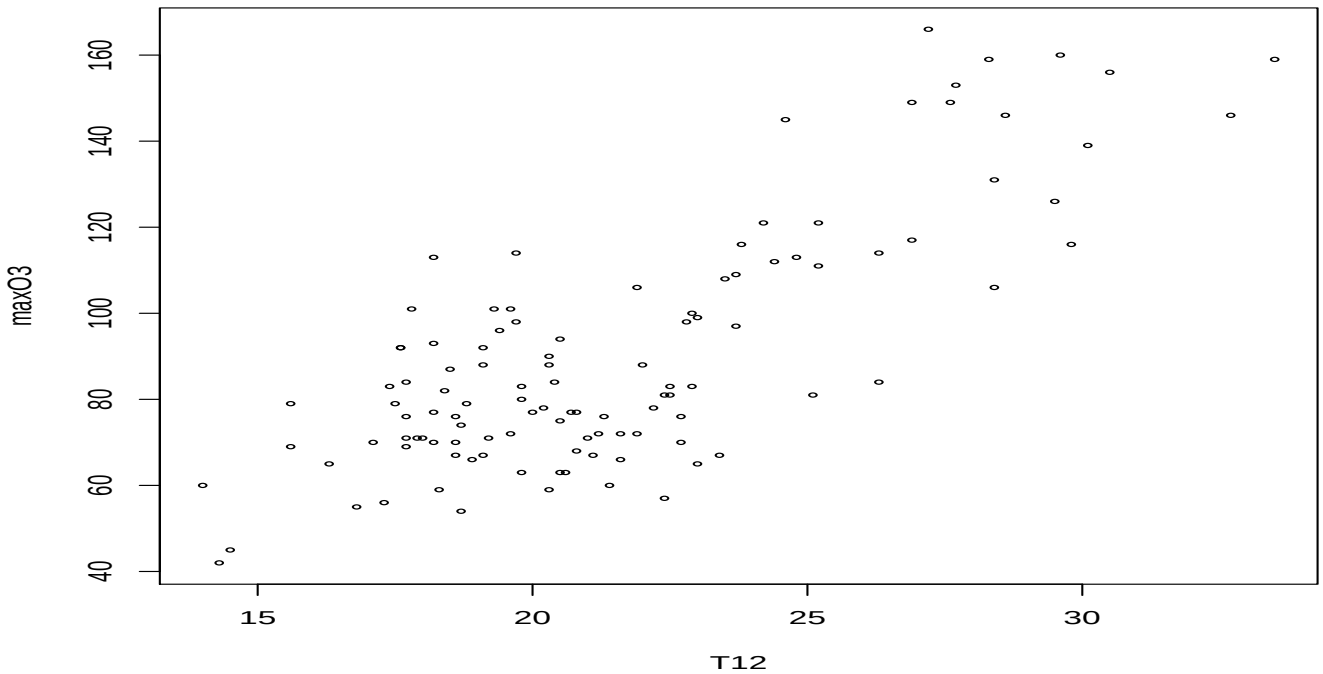
Comment expliquer (modéliser) la concentration en ozone à l'aide de toutes ces variables ?

Ozone en fonction de la température à 12h ?

MaxO3	87	82	92	114	94	80	...
T12	18.5	18.4	17.6	19.7	20.5	19.8	...

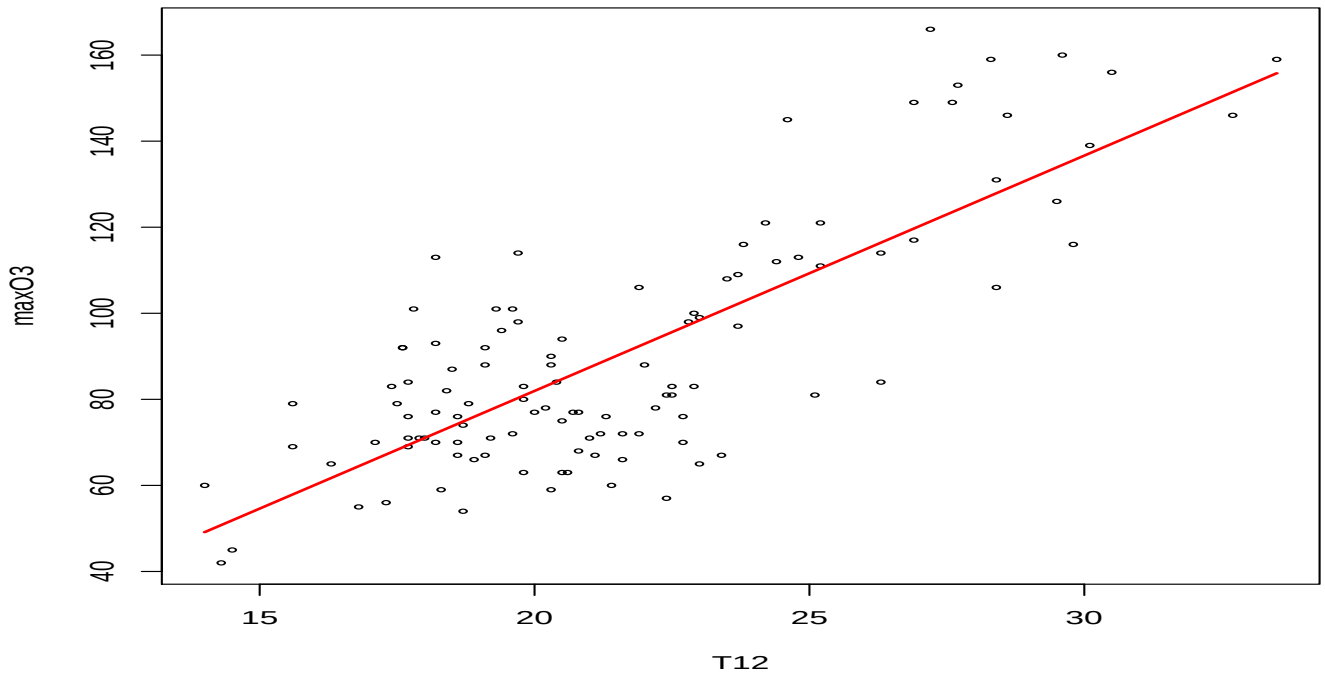
Ozone en fonction de la température à 12h ?

MaxO3	87	82	92	114	94	80	...
T12	18.5	18.4	17.6	19.7	20.5	19.8	...



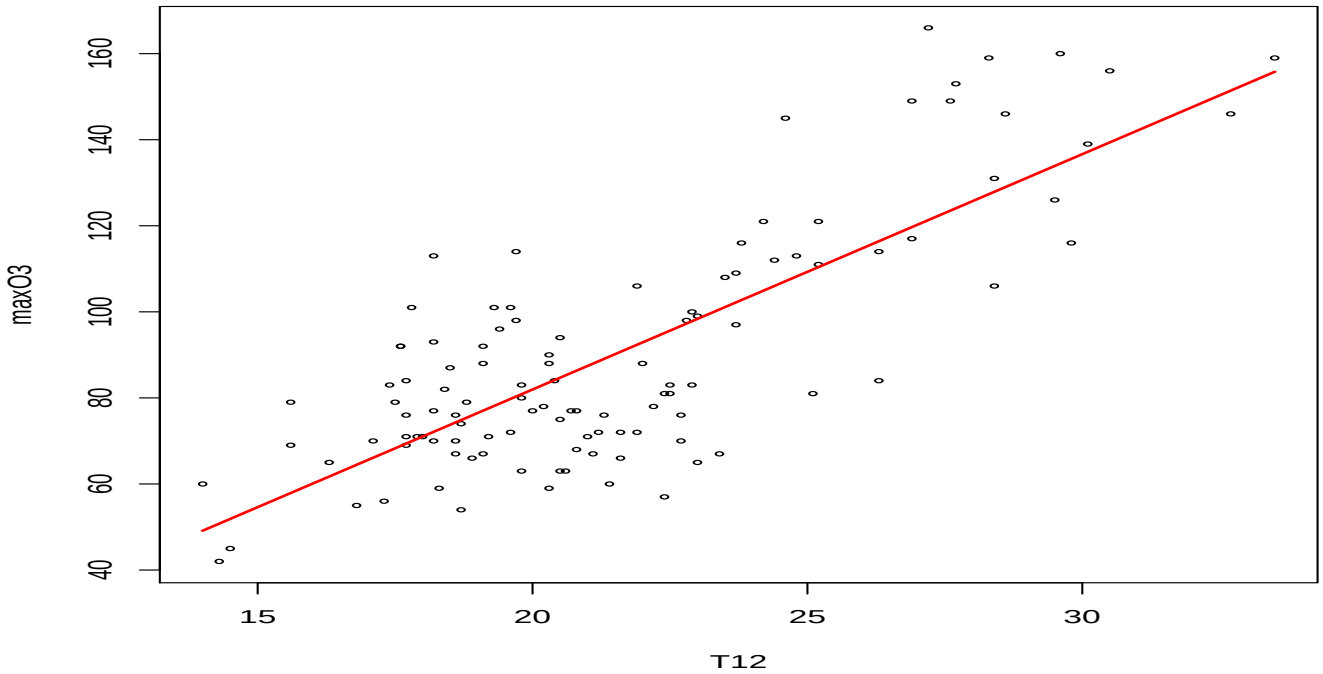
Ozone en fonction de la température à 12h ?

MaxO3	87	82	92	114	94	80	...
T12	18.5	18.4	17.6	19.7	20.5	19.8	...



Ozone en fonction de la température à 12h ?

MaxO3	87	82	92	114	94	80	...
T12	18.5	18.4	17.6	19.7	20.5	19.8	...



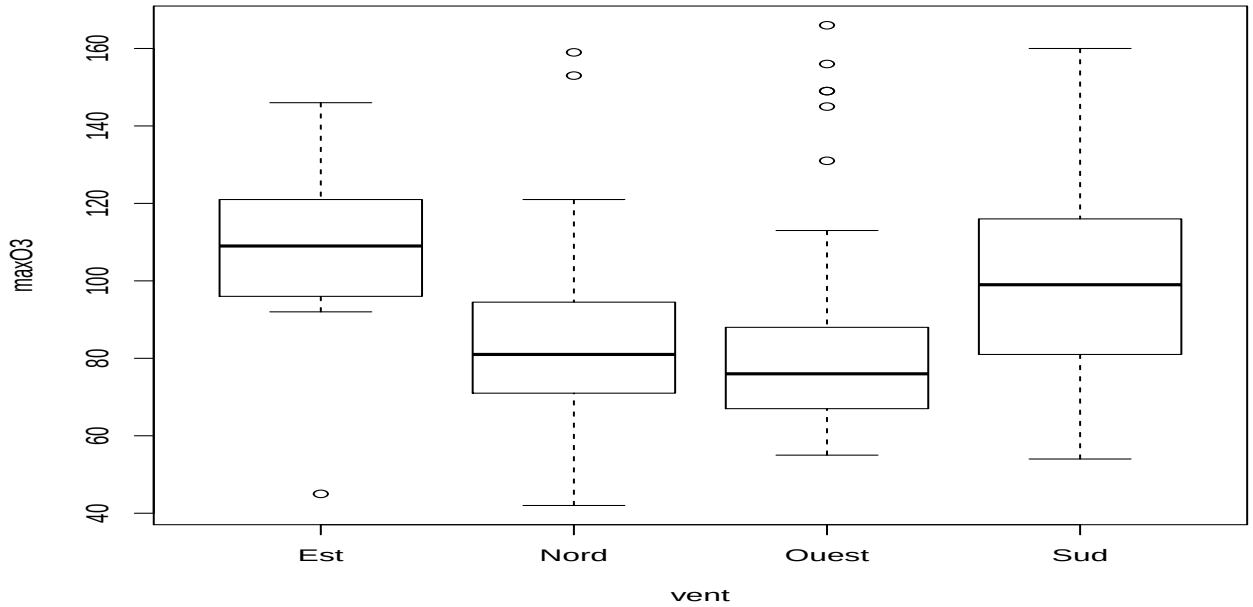
Comment ajuster le nuage de points ?

Ozone en fonction du vent ?

MaxO3	87	82	92	114	94	80	...
Vent	Nord	Nord	Est	Nord	Ouest	Ouest	...

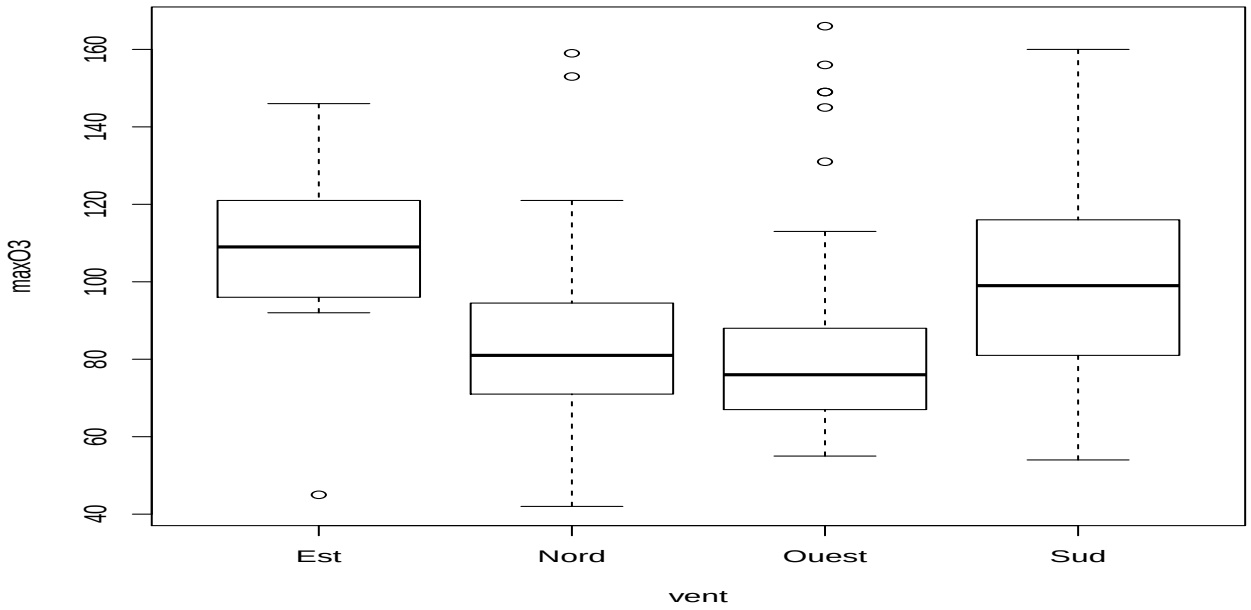
Ozone en fonction du vent ?

MaxO3	87	82	92	114	94	80	...
Vent	Nord	Nord	Est	Nord	Ouest	Ouest	...



Ozone en fonction du vent ?

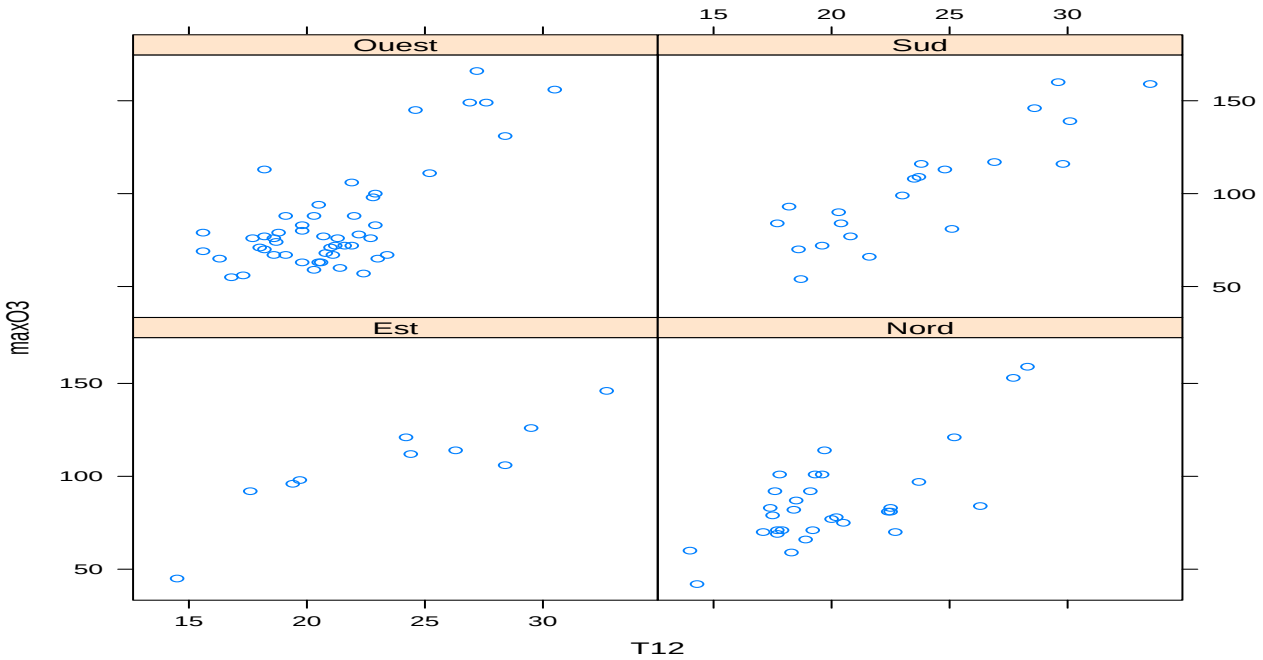
MaxO3	87	82	92	114	94	80	...
Vent	Nord	Nord	Est	Nord	Ouest	Ouest	...



$$MaxO3 \approx \alpha_1 \mathbf{1}_{Vent=est} + \ldots + \alpha_4 \mathbf{1}_{Vent=sud}.$$

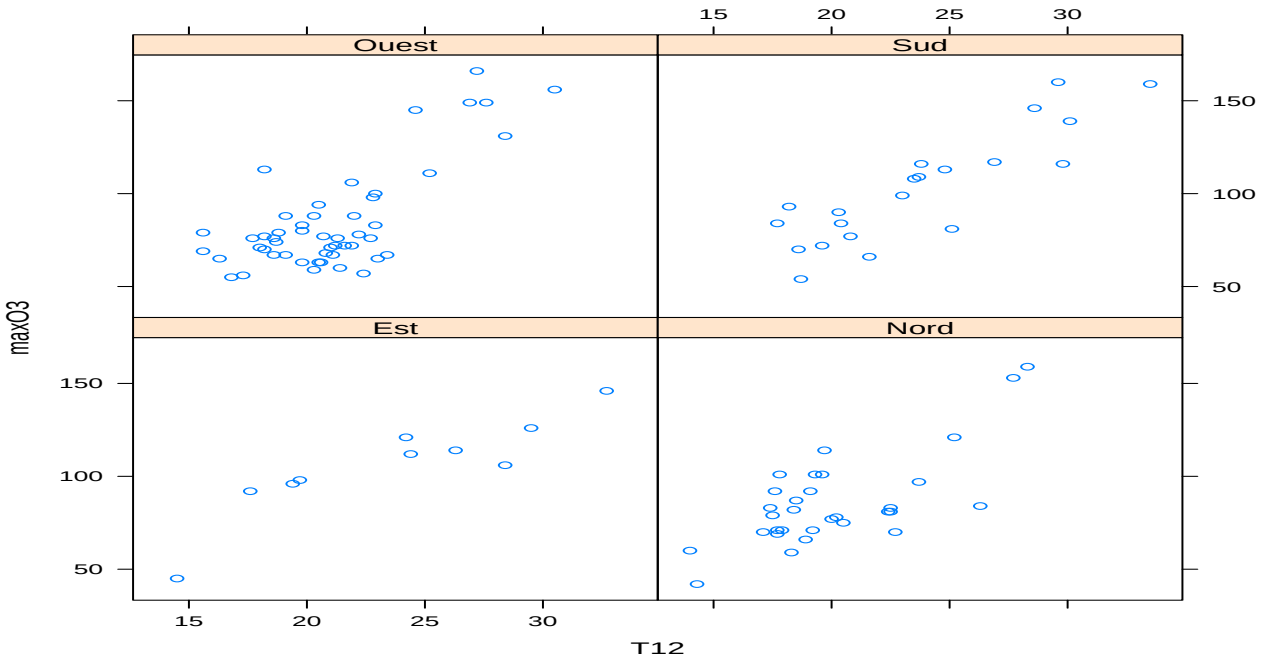
$$\alpha_j = ???$$

Ozone en fonction de la température à 12h et du vent ?



$$maxO3 \approx \begin{cases} \beta_{01} + \beta_{11} T12 & \text{si vent=est} \\ \vdots & \vdots \\ \beta_{04} + \beta_{14} T12 & \text{si vent=ouest} \end{cases}$$

Ozone en fonction de la température à 12h et du vent ?



$$maxO3 \approx \begin{cases} \beta_{01} + \beta_{11} T12 & \text{si vent=est} \\ \vdots & \vdots \\ \beta_{04} + \beta_{14} T12 & \text{si vent=ouest} \end{cases}$$

Autre modélisation

- Généralisation

$$maxO3 \approx \beta_0 + \beta_1 V_1 + \ldots + \beta_{12} V_{12}$$

Questions

- Comment calculer (ou plutôt **estimer**) les paramètres β_j ?
- Le modèle avec les 12 variables est-il "meilleur" que des modèles avec moins de variables ?
- Comment trouver le "meilleur" sous-groupe de variables ?

Autre modélisation

- Généralisation

$$maxO3 \approx \beta_0 + \beta_1 V_1 + \ldots + \beta_{12} V_{12}$$

Questions

- Comment calculer (ou plutôt **estimer**) les paramètres β_j ?
- Le modèle avec les 12 variables est-il "meilleur" que des modèles avec moins de variables ?
- Comment trouver le "meilleur" sous-groupe de variables ?

Notations

- Y : variable (aléatoire) à expliquer à valeurs dans $\mathbb{R}$.
- $X_1, \ldots, X_p$: p variables explicatives à valeurs dans $\mathbb{R}$.
- n observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ avec $x_i = (x_{i1}, \ldots, x_{ip})$.

Le modèle de régression linéaire multiple

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

où les erreurs aléatoires ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Notations

- Y : variable (aléatoire) à expliquer à valeurs dans $\mathbb{R}$.
- $X_1, \ldots, X_p$: p variables explicatives à valeurs dans $\mathbb{R}$.
- n observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ avec $x_i = (x_{i1}, \ldots, x_{ip})$.

Le modèle de régression linéaire multiple

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

où les erreurs aléatoires ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Ecriture matricielle

- On note

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ecriture matricielle

Le modèle se réécrit

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

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I_n})$.

Ecriture matricielle

- On note

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ecriture matricielle

Le modèle se réécrit

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

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I_n})$.

Estimateurs des moindres carrés

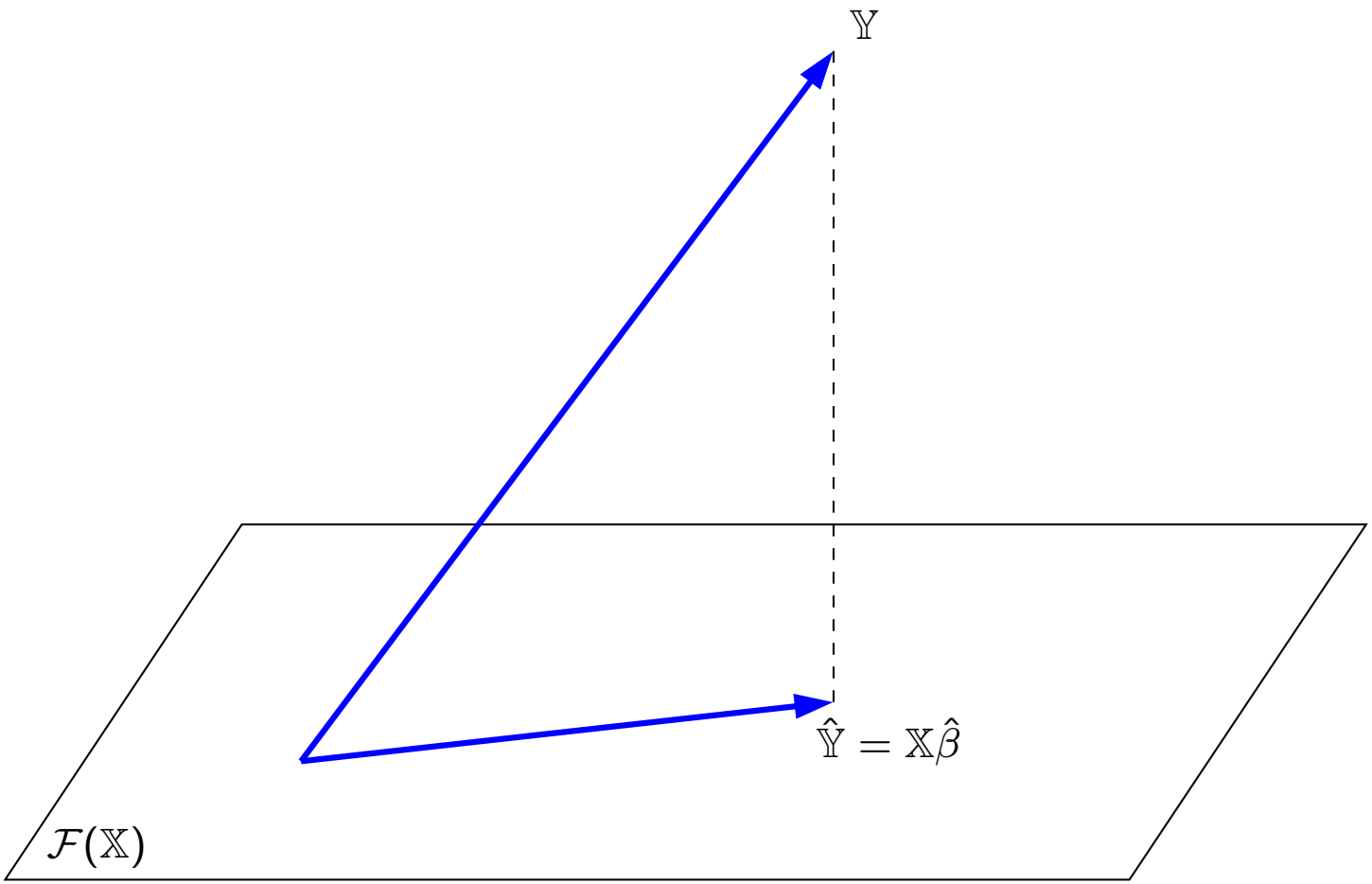
Définition

On appelle **estimateur des moindres carrés** $\hat{\beta}$ de β la statistique suivante :

$$\hat{\beta} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_p x_{ip})^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- On note $\mathcal{F}(\mathbb{X})$ le s.e.v. de $\mathbb{R}^n$ de dimension $p+1$ engendré par les $p+1$ colonnes de $\mathbb{X}$.
- Chercher l'estimateur des moindres carrés revient à minimiser la distance entre $\mathbb{Y} \in \mathbb{R}^n$ et $\mathcal{F}(\mathbb{X})$.

Représentation géométrique



Expression de l'estimateur des moindres carrés

- On déduit que $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de $\mathbb{Y}$ sur $\mathcal{F}(\mathbb{X})$:

$$\mathbb{X}\hat{\beta} = \mathbf{P}_{\mathcal{F}(\mathbb{X})}(\mathbb{Y}) = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Théorème

Si la matrice $\mathbb{X}$ est de plein rang, l'estimateur des MC est donné par :

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

Expression de l'estimateur des moindres carrés

- On déduit que $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de $\mathbb{Y}$ sur $\mathcal{F}(\mathbb{X})$:

$$\mathbb{X}\hat{\beta} = \mathbf{P}_{\mathcal{F}(\mathbb{X})}(\mathbb{Y}) = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Théorème

Si la matrice $\mathbb{X}$ est de plein rang, l'estimateur des MC est donné par :

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

Expression de l'estimateur des moindres carrés

- On déduit que $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de $\mathbb{Y}$ sur $\mathcal{F}(\mathbb{X})$:

$$\mathbb{X}\hat{\beta} = \mathbf{P}_{\mathcal{F}(\mathbb{X})}(\mathbb{Y}) = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Théorème

Si la matrice $\mathbb{X}$ est de plein rang, l'estimateur des MC est donné par :

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

Propriété

1. $\hat{\beta}$ est un estimateur sans biais de β.
2. La matrice de variance-covariance de $\hat{\beta}$ est donnée par

$$\mathbf{V}(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}.$$

3. $\hat{\beta}$ est VUMSB.

Loi des estimateurs

- Soit $\hat{\varepsilon} = \mathbb{Y} - \hat{\mathbb{Y}}$ le vecteur des résidus et $\widehat{\sigma^2}$ l'estimateur de σ^2 défini par

$$\widehat{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}.$$

Proposition

1. $\hat{\beta}$ est un vecteur gaussien d'espérance β et de matrice de variance-covariance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
2. $(n-(p+1))\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
3. $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendantes.

Loi des estimateurs

- Soit $\hat{\varepsilon} = \mathbb{Y} - \hat{\mathbb{Y}}$ le vecteur des résidus et $\widehat{\sigma^2}$ l'estimateur de σ^2 défini par

$$\widehat{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}.$$

Proposition

1. $\hat{\beta}$ est un vecteur gaussien d'espérance β et de matrice de variance-covariance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
2. $(n-(p+1))\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
3. $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendantes.

Loi des estimateurs

- Soit $\hat{\varepsilon} = \mathbb{Y} - \hat{\mathbb{Y}}$ le vecteur des résidus et $\widehat{\sigma^2}$ l'estimateur de σ^2 défini par

$$\widehat{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}.$$

Proposition

1. $\hat{\beta}$ est un vecteur gaussien d'espérance β et de matrice de variance-covariance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
2. $(n-(p+1))\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
3. $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendantes.

Corollaire

On note $\widehat{\sigma_j^2} = \widehat{\sigma^2}[\mathbb{X}'\mathbb{X}]_{jj}^{-1}$ pour $j = 0, \ldots, p$. On a

$$\forall j = 0, \ldots, p, \quad \frac{\hat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \sim \mathcal{T}(n - (p+1)).$$

On déduit de ce corollaire :

- des intervalles de confiance de niveau $1 - \alpha$ pour β_j.
- des procédures de test pour des hypothèses du genre $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Corollaire

On note $\widehat{\sigma_j^2} = \widehat{\sigma^2}[\mathbb{X}'\mathbb{X}]^{-1}_{jj}$ pour $j = 0, \ldots, p$. On a

$$\forall j = 0, \ldots, p, \quad \frac{\hat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \sim \mathcal{T}(n - (p+1)).$$

On déduit de ce corollaire :

- des intervalles de confiance de niveau $1 - \alpha$ pour β_j.
- des procédures de test pour des hypothèses du genre $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Prévision

- On dispose d'une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ et on souhaite prédire la valeur $y_{n+1} = x'_{n+1}\beta$ associée à cette nouvelle observation.

- Un estimateur (naturel) de y_{n+1} est $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$.
- Un intervalle de confiance de niveau $1-\alpha$ pour y_{n+1} est donné par

$$\left[\hat{y}_{n+1} \pm t_{n-(p+1)}(\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x_{n+1}+1}\right].$$

Prévision

- On dispose d'une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ et on souhaite prédire la valeur $y_{n+1} = x'_{n+1}\beta$ associée à cette nouvelle observation.

- Un estimateur (naturel) de y_{n+1} est $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$.
- Un intervalle de confiance de niveau $1-\alpha$ pour y_{n+1} est donné par

$$\left[\hat{y}_{n+1} \pm t_{n-(p+1)}(\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x_{n+1}+1}\right].$$

Prévision

- On dispose d'une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ et on souhaite prédire la valeur $y_{n+1} = x'_{n+1}\beta$ associée à cette nouvelle observation.

- Un estimateur (naturel) de y_{n+1} est $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$.
- Un intervalle de confiance de niveau $1-\alpha$ pour y_{n+1} est donné par

$$\left[\hat{y}_{n+1} \pm t_{n-(p+1)}(\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x_{n+1}+1}\right].$$

Exemple de l'ozone

- On considère le modèle de régression multiple :

$$MaxO3 = \beta_0 + \beta_1 T_{12} + \beta_2 T_{15} + \beta_3 N_{12} + \beta_4 V_{12} + \beta_5 MaxO3v + \varepsilon.$$

```
> reg.multi <- lm(maxO3~T12+T15+Ne12+Vx12+maxO3v,data=donnees)
> summary(reg.multi)

Call:
lm(formula = maxO3 ~ T12 + T15 + Ne12 + Vx12 + maxO3v, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-54.216  -9.446  -0.896   8.007  41.186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.04498   13.01591   0.234   0.8155
T12          2.47747    1.09257   2.268   0.0254 *
T15          0.63177    0.96382   0.655   0.5136
Ne12        -1.83560    0.89439  -2.052   0.0426 *
Vx12         1.33295    0.58168   2.292   0.0239 *
maxO3v       0.34215    0.05989   5.713 1.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.58 on 106 degrees of freedom
Multiple R-squared: 0.7444,Adjusted R-squared: 0.7324
F-statistic: 61.75 on 5 and 106 DF,  p-value: < 2.2e-16
```

"Validation" du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

"Validation" du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

"Validation" du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

"Validation" du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

"Validation" du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

Une autre écriture du modèle linéaire

- Le modèle linéaire

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

- peut se réécrire pour $i = 1, \ldots, n$

$$\mathcal{L}(Y_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

Interprétation

Au point x_i la loi de Y est une gaussienne $\mathcal{N}(x_i'\beta, \sigma^2)$.

Une autre écriture du modèle linéaire

- Le modèle linéaire

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

- peut se réécrire pour $i = 1, \ldots, n$

$$\mathcal{L}(Y_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

Interprétation

Au point x_i la loi de Y est une gaussienne $\mathcal{N}(x_i'\beta, \sigma^2)$.

Une autre écriture du modèle linéaire

- Le modèle linéaire

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

- peut se réécrire pour $i = 1, \ldots, n$

$$\mathcal{L}(Y_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

Interprétation

Au point x_i la loi de Y est une gaussienne $\mathcal{N}(x_i'\beta, \sigma^2)$.

- On peut alors calculer la (log)-vraisemblance du modèle

$$\mathcal{L}(y_1, \ldots, y_n; \beta) = \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- **Conclusion :** l'estimateur du maximum de vraisemblance $\hat{\beta}_{MV}$ coïncide avec l'estimateur des moindres carrés $\hat{\beta}$.

Remarque

- Si les variable explicatives sont aléatoires, ce n'est plus la loi de Y_i qui est modélisée mais celle de Y_i sachant $X_i = x_i$

$$\mathcal{L}(Y_i | X_i = x_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

- Plus généralement, lorsque les variables explicatives sont supposées aléatoires (économétrie), poser un modèle de régression revient à "mettre" une famille de loi sur Y sachant $X = x$.

- On peut alors calculer la (log)-vraisemblance du modèle

$$\mathcal{L}(y_1, \ldots, y_n; \beta) = \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- **Conclusion :** l'estimateur du maximum de vraisemblance $\hat{\beta}_{MV}$ coïncide avec l'estimateur des moindres carrés $\hat{\beta}$.

Remarque

- Si les variable explicatives sont aléatoires, ce n'est plus la loi de Y_i qui est modélisée mais celle de Y_i sachant $X_i = x_i$

$$\mathcal{L}(Y_i | X_i = x_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

- Plus généralement, lorsque les variables explicatives sont supposées aléatoires (économétrie), poser un modèle de régression revient à "mettre" une famille de loi sur Y sachant $X = x$.

- On peut alors calculer la (log)-vraisemblance du modèle

$$\mathcal{L}(y_1, \ldots, y_n; \beta) = \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- **Conclusion :** l'estimateur du maximum de vraisemblance $\hat{\beta}_{MV}$ coïncide avec l'estimateur des moindres carrés $\hat{\beta}$.

Remarque

- Si les variable explicatives sont aléatoires, ce n'est plus la loi de Y_i qui est modélisée mais celle de Y_i sachant $X_i = x_i$

$$\mathcal{L}(Y_i | X_i = x_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

- Plus généralement, lorsque les variables explicatives sont supposées aléatoires (économétrie), poser un modèle de régression revient à "mettre" une famille de loi sur Y sachant $X = x$.

Détection de clients à risque

- Une chaine de magasin a mis en place une carte de crédit.
- Elle dispose d'un historique de 145 clients dont 40 ont connu des défauts de paiement.
- Elle connait également d'autres caractéristiques de ces clients (sexe, taux d'enttement, revenus mensuels, dépenses effectuées sur certaines gammes de produit...)

Question

Comment prédire si un nouveau client connaîtra des défauts de paiement ?

Détection de clients à risque

- Une chaine de magasin a mis en place une carte de crédit.
- Elle dispose d'un historique de 145 clients dont 40 ont connu des défauts de paiement.
- Elle connait également d'autres caractéristiques de ces clients (sexe, taux d'enttement, revenus mensuels, dépenses effectuées sur certaines gammes de produit...)

Question

Comment prédire si un nouveau client connaîtra des défauts de paiement ?

Iris de Fisher

- On a mesuré sur 150 iris de 3 espèces différentes (Setosa, Versicolor, Virginica) les quantités suivantes :
  - Longueur et largeur des pétales
  - Longueur et largeur des sépales

Question

Comment identifier l'espèce d'un iris à partir de ces 4 caractéristiques ?

Iris de Fisher

- On a mesuré sur 150 iris de 3 espèces différentes (Setosa, Versicolor, Virginica) les quantités suivantes :
  - Longueur et largeur des pétales
  - Longueur et largeur des sépales

Question

Comment identifier l'espèce d'un iris à partir de ces 4 caractéristiques ?

Détection de spam

- Sur 4 601 mails, on a pu identifier 1813 spams.
- On a également mesuré sur chacun de ces mails la présence ou absence de 57 mots.

Question

Peut-on construire à partir de ces données une méthode de détection automatique de spam ?

Détection de spam

- Sur 4 601 mails, on a pu identifier 1813 spams.
- On a également mesuré sur chacun de ces mails la présence ou absence de 57 mots.

Question

Peut-on construire à partir de ces données une méthode de détection automatique de spam ?

Pathologie concernant les artères coronaires

- Problème : étudier la présence d'une pathologie concernant les artères coronaires en fonction de l'âge des individus.

- Données : on dispose d'un échantillon de taille 100 sur lequel on a mesuré les variables :
  - chd qui vaut 1 si la pathologie est présente, 0 sinon ;
  - age qui correspond à l'âge de l'individu.

```
> artere[1:5,]
   age agrp chd
1.  20    1   0
2.  23    1   0
3.  24    1   0
4.  25    1   0
5.  25    1   1
```

Pathologie concernant les artères coronaires

- Problème : étudier la présence d'une pathologie concernant les artères coronaires en fonction de l'âge des individus.

- Données : on dispose d'un échantillon de taille 100 sur lequel on a mesuré les variables :
  - chd qui vaut 1 si la pathologie est présente, 0 sinon ;
  - age qui correspond à l'âge de l'individu.

```
> artere[1:5,]
   age agrp chd
1.  20    1   0
2.  23    1   0
3.  24    1   0
4.  25    1   0
5.  25    1   1
```

Représentation du problème

- Tous ces problèmes peuvent être appréhendés dans un contexte de régression : on cherche à expliquer une variable Y par d'autres variables $X_1, \ldots, X_p$:

Y	X
Défaut de paiement	caractéristiques du client
Espèce de l'iris	Longueur, largeur pétales et sépales
Spam	présence/absence de mots

- La variable à expliquer n'est plus quantitative mais qualitative.
- On parle de problème de discrimination ou classification supervisée.

Représentation du problème

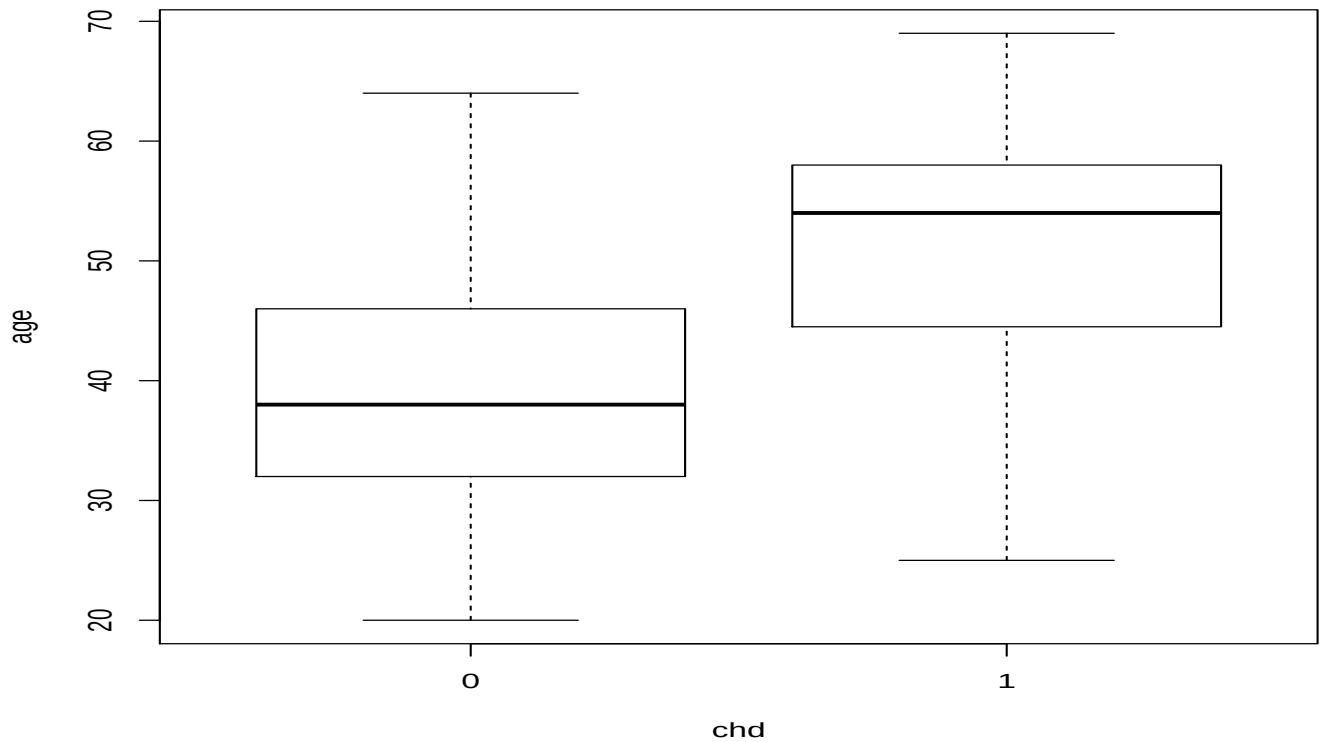
- Tous ces problèmes peuvent être appréhendés dans un contexte de régression : on cherche à expliquer une variable Y par d'autres variables $X_1, \ldots, X_p$:

Y	X
Défaut de paiement	caractéristiques du client
Espèce de l'iris	Longueur, largeur pétales et sépales
Spam	présence/absence de mots

- La variable à expliquer n'est plus quantitative mais qualitative.
- On parle de problème de discrimination ou classification supervisée.

Boxplot

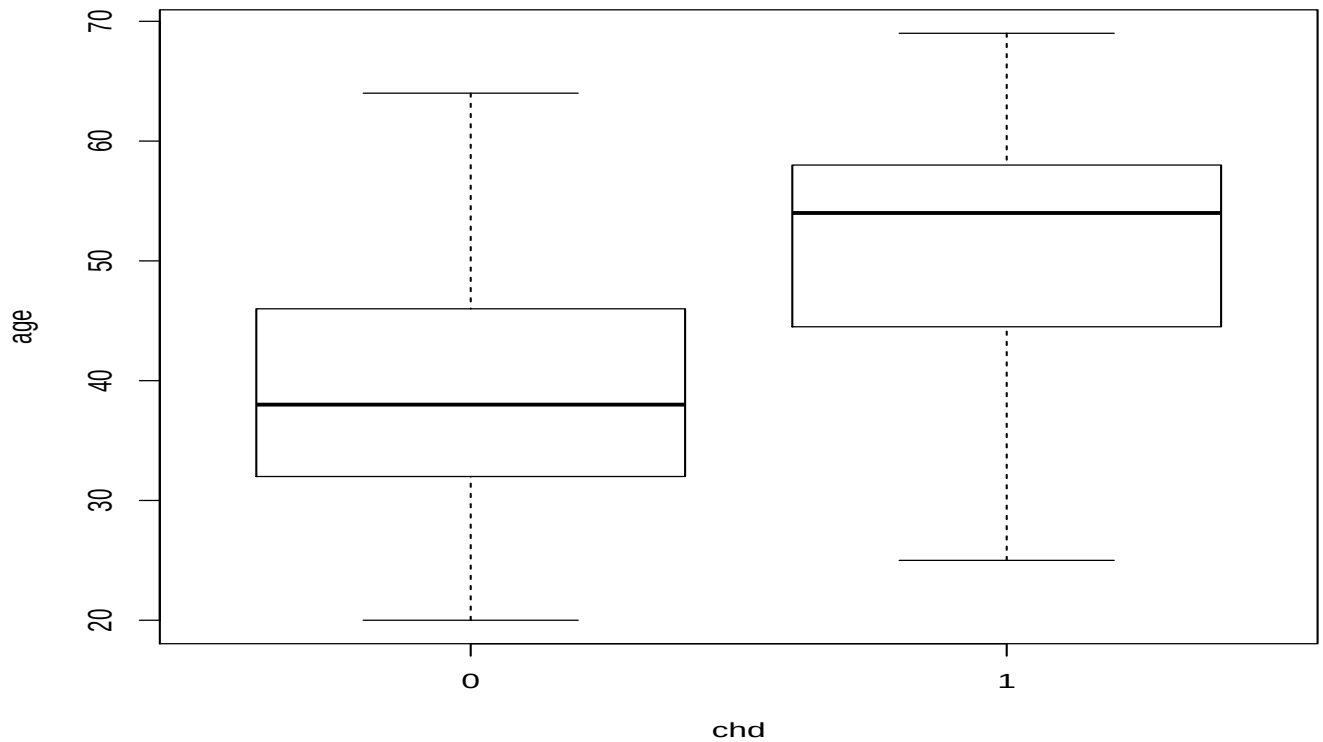
```
> plot(age~chd,data=artere)
```



Il semble que la maladie a plus de chance d'être présente chez les personnes agées.

Boxplot

```
> plot(age~chd,data=artere)
```



Il semble que la maladie a plus de chance d'être présente chez les personnes agées.

Début de modélisation

Question

Comment expliquer la relation entre la maladie et l'âge ?

- On désigne par
  - Y la variable aléatoire qui prend pour valeur 1 si l'individu est atteint, 0 sinon.
  - X la variable (aléatoire) qui correspond à l'âge de l'individu.

Le problème consiste ainsi à tenter de quantifier la relation entre Y et X à partir des données, c'est-à-dire d'un échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 100$.

Début de modélisation

Question

Comment expliquer la relation entre la maladie et l'âge ?

- On désigne par
  - Y la variable aléatoire qui prend pour valeur 1 si l'individu est atteint, 0 sinon.
  - X la variable (aléatoire) qui correspond à l'âge de l'individu.

Le problème consiste ainsi à tenter de quantifier la relation entre Y et X à partir des données, c'est-à-dire d'un échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 100$.

Début de modélisation

Question

Comment expliquer la relation entre la maladie et l'âge ?

- On désigne par
  - Y la variable aléatoire qui prend pour valeur 1 si l'individu est atteint, 0 sinon.
  - X la variable (aléatoire) qui correspond à l'âge de l'individu.

Le problème consiste ainsi à tenter de quantifier la relation entre Y et X à partir des données, c'est-à-dire d'un échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 100$.

Première idée

- On se base sur le modèle linéaire.

- On suppose que les deux variables Y et X sont liées par une relation de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

où $\beta_0 \in \mathbb{R}$ et $\beta_1 \in \mathbb{R}$ sont les paramètres inconnus du modèle et ε est une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$.

Problème

La variable Y est ici **qualitative**, l'écriture (1) n'a donc aucun sens.
$\Longrightarrow$ mauvaise idée

Première idée

- On se base sur le modèle linéaire.
- On suppose que les deux variables Y et X sont liées par une relation de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

où $\beta_0 \in \mathbb{R}$ et $\beta_1 \in \mathbb{R}$ sont les paramètres inconnus du modèle et ε est une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$.

Problème

La variable Y est ici **qualitative**, l'écriture (1) n'a donc aucun sens.
$\Longrightarrow$ mauvaise idée

Première idée

- On se base sur le modèle linéaire.
- On suppose que les deux variables Y et X sont liées par une relation de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

où $\beta_0 \in \mathbb{R}$ et $\beta_1 \in \mathbb{R}$ sont les paramètres inconnus du modèle et ε est une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$.

Problème

La variable Y est ici **qualitative**, l'écriture (1) n'a donc aucun sens.
$\Longrightarrow$ mauvaise idée

Loi conditionnelle

- Chercher à expliquer Y par X revient à chercher de l'information sur la loi de probabilité de Y sachant X.

- En effet, le modèle de régression linéaire peut se réécrire en caractérisant la loi de $Y|X = x$ par la loi $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Idée

- Etendre cette caractérisation à notre contexte (où la variable à expliquer est binaire).
- Une loi candidate naturelle pour la variable $Y|X = x$ est la loi de Bernoulli.

Loi conditionnelle

- Chercher à expliquer Y par X revient à chercher de l'information sur la loi de probabilité de Y sachant X.

- En effet, le modèle de régression linéaire peut se réécrire en caractérisant la loi de $Y|X = x$ par la loi $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Idée

- Etendre cette caractérisation à notre contexte (où la variable à expliquer est binaire).
- Une loi candidate naturelle pour la variable $Y|X = x$ est la loi de Bernoulli.

Loi conditionnelle

- Chercher à expliquer Y par X revient à chercher de l'information sur la loi de probabilité de Y sachant X.

- En effet, le modèle de régression linéaire peut se réécrire en caractérisant la loi de $Y|X = x$ par la loi $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Idée

- Etendre cette caractérisation à notre contexte (où la variable à expliquer est binaire).
- Une loi candidate naturelle pour la variable $Y|X = x$ est la loi de Bernoulli.

Loi conditionnelle

- Chercher à expliquer Y par X revient à chercher de l'information sur la loi de probabilité de Y sachant X.

- En effet, le modèle de régression linéaire peut se réécrire en caractérisant la loi de $Y|X = x$ par la loi $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Idée

- Etendre cette caractérisation à notre contexte (où la variable à expliquer est binaire).
- Une loi candidate naturelle pour la variable $Y|X = x$ est la loi de Bernoulli.

Loi de Bernoulli

- On va ainsi caractériser la loi de $Y|X = x$ par la loi de Bernoulli.
- Cette loi dépend d'un paramètre

$$p(x) = \mathbf{P}(Y = 1|X = x).$$

- Sachant $X = x$, on a donc

$$Y = \begin{cases} 1 & \text{avec probabilité } p(x) \\ 0 & \text{avec probabilité } 1 - p(x) \end{cases}$$

La modélisation

Il reste maintenant à caractériser la probabilité $p(x)$.

Loi de Bernoulli

- On va ainsi caractériser la loi de $Y|X = x$ par la loi de Bernoulli.
- Cette loi dépend d'un paramètre

$$p(x) = \mathbf{P}(Y = 1|X = x).$$

- Sachant $X = x$, on a donc

$$Y = \begin{cases} 1 & \text{avec probabilité } p(x) \\ 0 & \text{avec probabilité } 1 - p(x) \end{cases}$$

La modélisation

Il reste maintenant à caractériser la probabilité $p(x)$.

Loi de Bernoulli

- On va ainsi caractériser la loi de $Y|X = x$ par la loi de Bernoulli.
- Cette loi dépend d'un paramètre

$$p(x) = \mathbf{P}(Y = 1|X = x).$$

- Sachant $X = x$, on a donc

$$Y = \begin{cases} 1 & \text{avec probabilité } p(x) \\ 0 & \text{avec probabilité } 1 - p(x) \end{cases}$$

La modélisation

Il reste maintenant à caractériser la probabilité $p(x)$.

Loi de Bernoulli

- On va ainsi caractériser la loi de $Y|X = x$ par la loi de Bernoulli.
- Cette loi dépend d'un paramètre

$$p(x) = \mathbf{P}(Y = 1|X = x).$$

- Sachant $X = x$, on a donc

$$Y = \begin{cases} 1 & \text{avec probabilité } p(x) \\ 0 & \text{avec probabilité } 1 - p(x) \end{cases}$$

La modélisation

Il reste maintenant à caractériser la probabilité $p(x)$.

Première idée

- Là encore, on peut se baser sur le modèle linéaire et proposer

$$p(x) = \beta_0 + \beta_1 x.$$

- Cette écriture n'est pas satisfaisante. En effet
  - $p(x) \in [0, 1]$ tandis que $\beta_0 + \beta_1 x \in \mathbb{R}$.
  - Idée : trouver une transformation φ de $p(x)$ telle que $\varphi(p(x))$ prenne ses valeurs dans $\mathbb{R}$.

Première idée

- Là encore, on peut se baser sur le modèle linéaire et proposer

$$p(x) = \beta_0 + \beta_1 x.$$

- Cette écriture n'est pas satisfaisante. En effet
  - $p(x) \in [0, 1]$ tandis que $\beta_0 + \beta_1 x \in \mathbb{R}$.
  - Idée : trouver une transformation φ de $p(x)$ telle que $\varphi(p(x))$ prenne ses valeurs dans $\mathbb{R}$.

Première idée

- Là encore, on peut se baser sur le modèle linéaire et proposer

$$p(x) = \beta_0 + \beta_1 x.$$

- Cette écriture n'est pas satisfaisante. En effet
  - $p(x) \in [0, 1]$ tandis que $\beta_0 + \beta_1 x \in \mathbb{R}$.
  - Idée : trouver une transformation φ de $p(x)$ telle que $\varphi(p(x))$ prenne ses valeurs dans $\mathbb{R}$.

Première idée

- Là encore, on peut se baser sur le modèle linéaire et proposer

$$p(x) = \beta_0 + \beta_1 x.$$

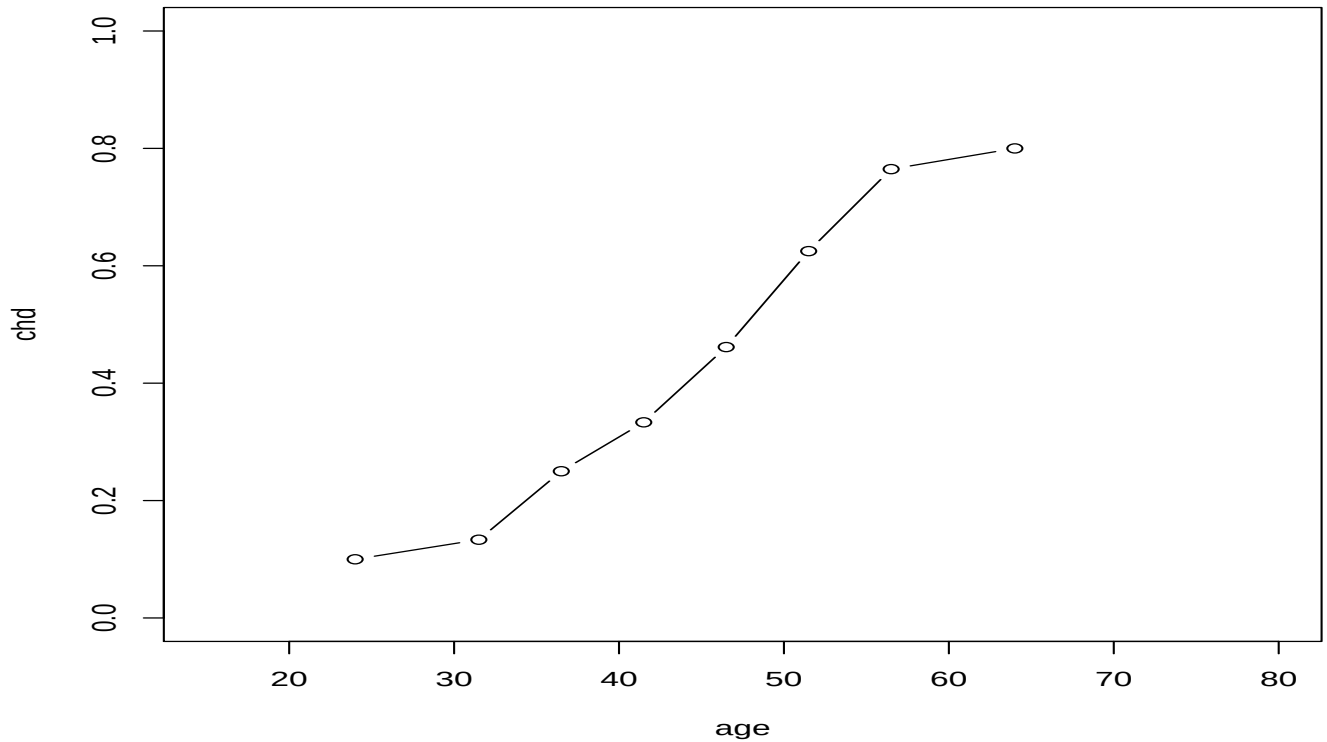
- Cette écriture n'est pas satisfaisante. En effet
  - $p(x) \in [0, 1]$ tandis que $\beta_0 + \beta_1 x \in \mathbb{R}$.
  - Idée : trouver une transformation φ de $p(x)$ telle que $\varphi(p(x))$ prenne ses valeurs dans $\mathbb{R}$.

Transformation de $p(x)$

- On revient sur l'exemple du chd et on représente les fréquences cumulées d'apparition de la maladie en fonction de l'âge :

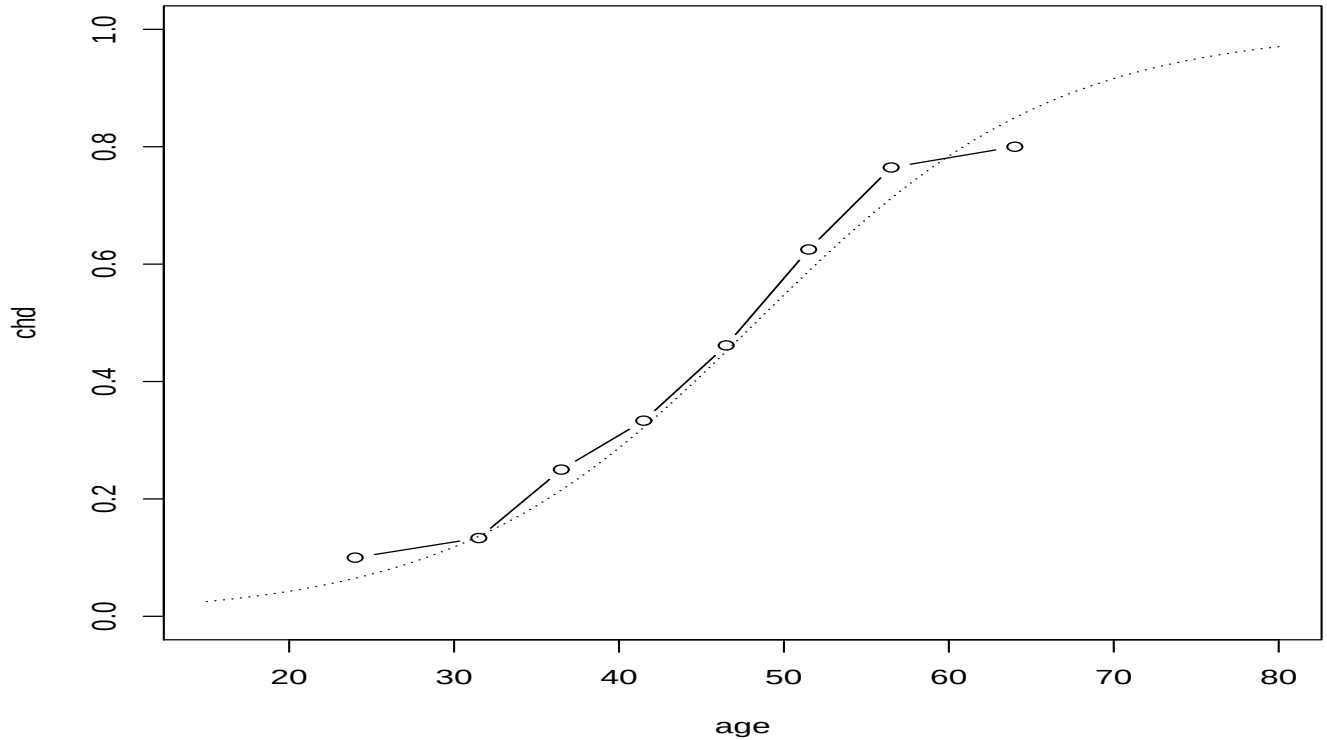
Transformation de $p(x)$

- On revient sur l'exemple du chd et on représente les fréquences cumulées d'apparition de la maladie en fonction de l'âge :



Transformation de $p(x)$

- On revient sur l'exemple du chd et on représente les fréquences cumulées d'apparition de la maladie en fonction de l'âge :



Trouver une transformation de $p(x)$ qui ajuste ce nuage de points.

Le modèle de régression logistique

- Il propose de modéliser la probabilité $p(x)$ selon

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

- On peut réécrire

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Le modèle de régression logistique

Le modèle de régression logistique consiste donc à caractériser la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p(x)$ tel que

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Le modèle de régression logistique

- Il propose de modéliser la probabilité $p(x)$ selon

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

- On peut réécrire

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Le modèle de régression logistique

Le modèle de régression logistique consiste donc à caractériser la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p(x)$ tel que

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Exemple sur R

```
> model <- glm(chd~age,data=artere,family=binomial)
> model

Call:  glm(formula = chd ~ age, family = binomial, data = artere)

Coefficients:
(Intercept)          age
    -5.3095       0.1109

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     136.7
Residual Deviance: 107.4  AIC: 111.4
```

- La fonction glm renvoie les estimations de β_0 et β_1.
- On peut ainsi avoir une estimation de la probabilité d'avoir une maladie pour un individu de 30 ans :

$$\hat{p}(x = 30) = \frac{\exp(-5.3095 + 0.1109 * 30)}{1 + \exp(-5.3095 + 0.1109 * 30)} \approx 0.12.$$

Exemple sur R

```
> model <- glm(chd~age,data=artere,family=binomial)
> model

Call:  glm(formula = chd ~ age, family = binomial, data = artere)

Coefficients:
(Intercept)          age
    -5.3095       0.1109

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     136.7
Residual Deviance: 107.4  AIC: 111.4
```

- La fonction glm renvoie les estimations de β_0 et β_1.
- On peut ainsi avoir une estimation de la probabilité d'avoir une maladie pour un individu de 30 ans :

$$\hat{p}(x = 30) = \frac{\exp(-5.3095 + 0.1109 * 30)}{1 + \exp(-5.3095 + 0.1109 * 30)} \approx 0.12.$$

Exemple sur R

```
> model <- glm(chd~age,data=artere,family=binomial)
> model

Call:  glm(formula = chd ~ age, family = binomial, data = artere)

Coefficients:
(Intercept)          age
    -5.3095       0.1109

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     136.7
Residual Deviance: 107.4  AIC: 111.4
```

- La fonction glm renvoie les estimations de β_0 et β_1.
- On peut ainsi avoir une estimation de la probabilité d'avoir une maladie pour un individu de 30 ans :

$$\hat{p}(x = 30) = \frac{\exp(-5.3095 + 0.1109 * 30)}{1 + \exp(-5.3095 + 0.1109 * 30)} \approx 0.12.$$

Exemple sur R

```
> model <- glm(chd~age,data=artere,family=binomial)
> model

Call:  glm(formula = chd ~ age, family = binomial, data = artere)

Coefficients:
(Intercept)          age
    -5.3095       0.1109

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     136.7
Residual Deviance: 107.4  AIC: 111.4
```

- La fonction glm renvoie les estimations de β_0 et β_1.
- On peut ainsi avoir une estimation de la probabilité d'avoir une maladie pour un individu de 30 ans :

$$\hat{p}(x = 30) = \frac{\exp(-5.3095 + 0.1109 * 30)}{1 + \exp(-5.3095 + 0.1109 * 30)} \approx 0.12.$$

1. Modèle statistique
   - Modèle de densité
   - Modèle de régression
   - Rappels sur le modèle de régression linéaire

2. Introduction au modèle de régression logistique
   - Exemples
   - Régression logistique simple

3. Le modèle linéaire généralisé
   - Introduction
   - Définitions
   - Modèle de Poisson

Le modèle logistique est un glm

- Le modèle de régression logistique s'ajuste sur R avec la fonction glm.
- Le modèle de régression logistique appartient à la famille des modèles linéaires généralisés.
- C'est pourquoi il faut spécifier l'argument `family=binomial` lorsque l'on veut faire une régression logistique.

Le modèle logistique est un glm

- Le modèle de régression logistique s'ajuste sur R avec la fonction glm.
- Le modèle de régression logistique appartient à la famille des modèles linéaires généralisés.
- C'est pourquoi il faut spécifier l'argument `family=binomial` lorsque l'on veut faire une régression logistique.

Le modèle logistique est un glm

- Le modèle de régression logistique s'ajuste sur R avec la fonction glm.
- Le modèle de régression logistique appartient à la famille des modèles linéaires généralisés.
- C'est pourquoi il faut spécifier l'argument `family=binomial` lorsque l'on veut faire une régression logistique.

Le modèle linéaire est un GLM

- Le modèle de régression linéaire s'ajuste sur R avec la fonction lm :

```
> Y <- rnorm(50)
> X <- runif(50)
> lm(Y~X)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

- Mais aussi avec la fonction glm :

```
> glm(Y~X,family=gaussian)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

Conclusion

Le modèle linéaire appartient également à la famille des modèles linéaires généralisés.

Le modèle linéaire est un GLM

- Le modèle de régression linéaire s'ajuste sur R avec la fonction lm :

```
> Y <- rnorm(50)
> X <- runif(50)
> lm(Y~X)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

- Mais aussi avec la fonction glm :

```
> glm(Y~X,family=gaussian)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

Conclusion

Le modèle linéaire appartient également à la famille des modèles linéaires généralisés.

Le modèle linéaire est un GLM

- Le modèle de régression linéaire s'ajuste sur R avec la fonction lm :

```
> Y <- rnorm(50)
> X <- runif(50)
> lm(Y~X)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

- Mais aussi avec la fonction glm :

```
> glm(Y~X,family=gaussian)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

Conclusion

Le modèle linéaire appartient également à la famille des modèles linéaires généralisés.

2 étapes identiques

- Les modèles linéaires et logistiques sont construits selon le même protocole en 2 étapes :

  1. Choix de la loi conditionnelle de $Y|X = x$:
     - Gaussienne pour le modèle linéaire ;
     - Bernoulli pour le modèle logistique.
  2. Choix d'une transformation g de l'espérance conditionnelle $\mathbf{E}[Y|X = x]$:
     - Logistique

$$g(\mathbf{E}[Y|X = x]) = g(p(x)) = \text{logit}\, p(x) = x'\beta$$

     - Linéaire

$$g(\mathbf{E}[Y|X = x]) = x'\beta.$$

2 étapes identiques

- Les modèles linéaires et logistiques sont construits selon le même protocole en 2 étapes :

  1. Choix de la loi conditionnelle de $Y|X = x$:
     - Gaussienne pour le modèle linéaire ;
     - Bernoulli pour le modèle logistique.
  2. Choix d'une transformation g de l'espérance conditionnelle $\mathbf{E}[Y|X = x]$:
     - Logistique

$$g(\mathbf{E}[Y|X = x]) = g(p(x)) = \text{logit}\, p(x) = x'\beta$$

     - Linéaire

$$g(\mathbf{E}[Y|X = x]) = x'\beta.$$

2 étapes identiques

- Les modèles linéaires et logistiques sont construits selon le même protocole en 2 étapes :

  1. Choix de la loi conditionnelle de $Y|X = x$:
     - Gaussienne pour le modèle linéaire ;
     - Bernoulli pour le modèle logistique.
  2. Choix d'une transformation g de l'espérance conditionnelle $\mathbf{E}[Y|X = x]$:
     - Logistique

$$g(\mathbf{E}[Y|X = x]) = g(p(x)) = \text{logit}\, p(x) = x'\beta$$

     - Linéaire

$$g(\mathbf{E}[Y|X = x]) = x'\beta.$$

Rappel : Famille exponentielle

Définition

Une loi de probabilité $\mathbf{P}$ appartient à une famille de lois de type exponentielle $\{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}^p}$ si il existe une mesure dominant μ (Lebesgue ou mesure de comptage le plus souvent) telle que les lois $\mathbf{P}_\theta$ admettent pour densité par rapport à ν

$$f_\theta(y) = c(\theta)h(y)\exp\left(\sum_{j=1}^{p}\alpha_j(\theta)\,T_j(y)\right)$$

où $T_1, \ldots, T_p$ sont des fonctions réelles mesurables.

Exemple : loi de Bernoulli

La loi de Bernoulli de paramètre p admet pour densité (par rapport à la mesure de comptage)

$$f_p(y) = (1-p)\exp(y\log(p/(1-p))).$$

Rappel : Famille exponentielle

Définition

Une loi de probabilité $\mathbf{P}$ appartient à une famille de lois de type exponentielle $\{\mathbf{P}_\theta\}_{\theta \in \mathbb{R}^p}$ si il existe une mesure dominant μ (Lebesgue ou mesure de comptage le plus souvent) telle que les lois $\mathbf{P}_\theta$ admettent pour densité par rapport à ν

$$f_\theta(y) = c(\theta) h(y) \exp\left(\sum_{j=1}^{p} \alpha_j(\theta) T_j(y) \right)$$

où $T_1, \ldots, T_p$ sont des fonctions réelles mesurables.

Exemple : loi de Bernoulli

La loi de Bernoulli de paramètre p admet pour densité (par rapport à la mesure de comptage)

$$f_p(y) = (1-p) \exp(y \log(p/(1-p))).$$

Rappel : Famille exponentielle

Définition

Une loi de probabilité $\mathbf{P}$ appartient à une famille de lois de type exponentielle $\{\mathbf{P}_\theta\}_{\theta \in \mathbb{R}^p}$ si il existe une mesure dominant μ (Lebesgue ou mesure de comptage le plus souvent) telle que les lois $\mathbf{P}_\theta$ admettent pour densité par rapport à ν

$$f_\theta(y) = c(\theta) h(y) \exp\left(\sum_{j=1}^{p} \alpha_j(\theta) T_j(y) \right)$$

où $T_1, \ldots, T_p$ sont des fonctions réelles mesurables.

Exemple : loi de Bernoulli

La loi de Bernoulli de paramètre p admet pour densité (par rapport à la mesure de comptage)

$$f_p(y) = (1-p) \exp(y \log(p/(1-p))).$$

Cadre

- On se place dans un contexte de régression : on cherche à expliquer une variable Y par p variables explicatives $X_1, \ldots, X_p$.
- On dispose d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i = (x_{i1}, \ldots, x_{ip})$ sont supposées fixes et les Y_i sont des variables aléatoires réelles indépendantes.

Cadre

- On se place dans un contexte de régression : on cherche à expliquer une variable Y par p variables explicatives $X_1, \ldots, X_p$.
- On dispose d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i = (x_{i1}, \ldots, x_{ip})$ sont supposées fixes et les Y_i sont des variables aléatoires réelles indépendantes.

Modèle linéaire généralisé : GLM

Un modèle linéaire généralisé est constitué de 3 composantes :

1. Composante aléatoire : la loi de probabilité de la réponse Y_i appartient à la famille exponentielle et est de la forme

$$f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right)$$

où a, b et c sont des fonctions spécifiées en fonction du type de la famille exponentielle.

2. Composante déterministe :

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

et précise quels sont les prédicteurs (on peut y inclure des transformations des prédicteurs, des interactions...).

3. Lien : spécifie le lien entre les deux composantes, plus précisément le lien entre l'espérance de Y_i et la composante déterministe : $g(\mathbf{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ où g est une fonction inversible appelée fonction de lien.

Modèle linéaire généralisé : GLM

Un modèle linéaire généralisé est constitué de 3 composantes :

1. Composante aléatoire : la loi de probabilité de la réponse Y_i appartient à la famille exponentielle et est de la forme

$$f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right)$$

où a, b et c sont des fonctions spécifiées en fonction du type de la famille exponentielle.

2. Composante déterministe :

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

et précise quels sont les prédicteurs (on peut y inclure des transformations des prédicteurs, des interactions...).

3. Lien : spécifie le lien entre les deux composantes, plus précisément le lien entre l'espérance de Y_i et la composante déterministe : $g(\mathbf{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ où g est une fonction inversible appelée fonction de lien.

Modèle linéaire généralisé : GLM

Un modèle linéaire généralisé est constitué de 3 composantes :

1. Composante aléatoire : la loi de probabilité de la réponse Y_i appartient à la famille exponentielle et est de la forme

$$f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right)$$

où a, b et c sont des fonctions spécifiées en fonction du type de la famille exponentielle.

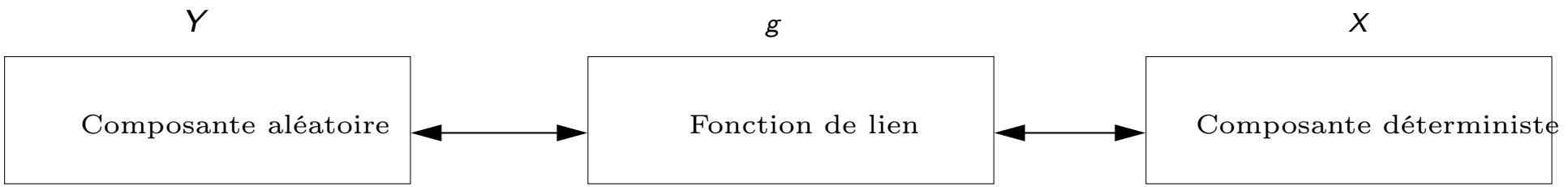
2. Composante déterministe :

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

et précise quels sont les prédicteurs (on peut y inclure des transformations des prédicteurs, des interactions...).

3. Lien : spécifie le lien entre les deux composantes, plus précisément le lien entre l'espérance de Y_i et la composante déterministe : $g(\mathbf{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ où g est une fonction inversible appelée fonction de lien.

Schéma GLM



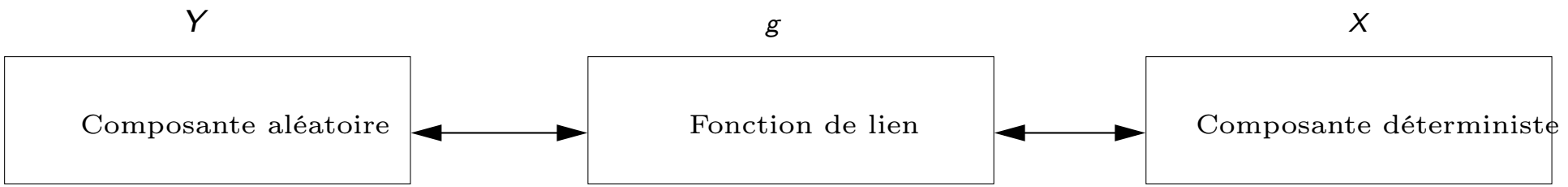
Y admet pour densité $f_\alpha(y)$

$\mathbf{E}[Y]$ est lié X selon $g(\mathbf{E}[Y]) = \eta(x)$

Choix de la combinaison linéaire des $\eta(x) = x'\beta$

Un modèle GLM sera caractérisé par le choix de ces trois composantes.

Schéma GLM



Y admet pour densité $f_\alpha(y)$

$\mathbf{E}[Y]$ est lié X selon $g(\mathbf{E}[Y]) = \eta(x)$

Choix de la combinaison linéaire des $\eta(x) = x'\beta$

Un modèle GLM sera caractérisé par le choix de ces trois composantes.

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Composante aléatoire et fonction de lien du modèle logistique

Propriété

Le modèle de régression logistique est un GLM.

En effet :

- La loi exponentielle est la loi de Bernoulli de paramètre $p_i = \mathbf{P}(Y_i = 1)$:

$$f_{\alpha_i}(y_i) = \exp[y_i x_i' \beta - \log(1 + \exp(x_i' \beta))].$$

On a donc $\alpha_i = x_i' \beta$ et $b(\alpha_i) = \log(1 + \exp(\alpha_i))$.

- La fonction de lien est

$$g(u) = \text{logit}(u) = \log \frac{u}{1-u}.$$

Composante aléatoire et fonction de lien du modèle logistique

Propriété

Le modèle de régression logistique est un GLM.

En effet :

- La loi exponentielle est la loi de Bernoulli de paramètre $p_i = \mathbf{P}(Y_i = 1)$:

$$f_{\alpha_i}(y_i) = \exp[y_i x_i' \beta - \log(1 + \exp(x_i' \beta))].$$

On a donc $\alpha_i = x_i' \beta$ et $b(\alpha_i) = \log(1 + \exp(\alpha_i))$.

- La fonction de lien est

$$g(u) = \text{logit}(u) = \log \frac{u}{1-u}.$$

Composante aléatoire et fonction de lien du modèle linéaire

Propriété

Le modèle linéaire gaussien est un GLM.

En effet :

- La loi exponentielle est la loi gaussienne de paramètres μ_i et σ^2 :

$$f_{\alpha_i}(y_i) = \exp\left\{\frac{y_i x_i'\beta - 0.5(x_i'\beta)^2}{\sigma^2} - \left(\frac{y_i^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\right)\right\}.$$

- La fonction de lien est l'identité.

Composante aléatoire et fonction de lien du modèle linéaire

Propriété

Le modèle linéaire gaussien est un GLM.

En effet :

- La loi exponentielle est la loi gaussienne de paramètres μ_i et σ^2 :

$$f_{\alpha_i}(y_i) = \exp\left\{\frac{y_i x_i'\beta - 0.5(x_i'\beta)^2}{\sigma^2} - \left(\frac{y_i^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\right)\right\}.$$

- La fonction de lien est l'identité.

Premier bilan

- Outre le choix classique de la composante déterministe (choix de la combinaison linéaire des variables explicatives), la modélisation GLM s'effectue à travers 2 choix :
  1. Choix de la loi de Y_i dans la famille exponentielle GLM décrite plus haut.
  2. Choix de la fonction de lien (inversible).

	logistique	linéaire
Loi expo	Bernoulli	Gaussienne
fdl	$g(u) = \text{logit}(u)$	$g(u) = u$

Premier bilan

- Outre le choix classique de la composante déterministe (choix de la combinaison linéaire des variables explicatives), la modélisation GLM s'effectue à travers 2 choix :

  1. Choix de la loi de Y_i dans la famille exponentielle GLM décrite plus haut.
  2. Choix de la fonction de lien (inversible).

	logistique	linéaire
Loi expo	Bernoulli	Gaussienne
fdl	$g(u) = \text{logit}(u)$	$g(u) = u$

Premier bilan

- Outre le choix classique de la composante déterministe (choix de la combinaison linéaire des variables explicatives), la modélisation GLM s'effectue à travers 2 choix :

  1. Choix de la loi de Y_i dans la famille exponentielle GLM décrite plus haut.
  2. Choix de la fonction de lien (inversible).

	logistique	linéaire
Loi expo	Bernoulli	Gaussienne
fdl	$g(u) = \text{logit}(u)$	$g(u) = u$

Premier bilan

- Outre le choix classique de la composante déterministe (choix de la combinaison linéaire des variables explicatives), la modélisation GLM s'effectue à travers 2 choix :

  1. Choix de la loi de Y_i dans la famille exponentielle GLM décrite plus haut.
  2. Choix de la fonction de lien (inversible).

	logistique	linéaire
Loi expo	Bernoulli	Gaussienne
fdl	$g(u) = \text{logit}(u)$	$g(u) = u$

Choix de la loi exponentielle et de la fonction de lien

1. Loi exponentielle. Ce choix est généralement guidé par la nature de la variable à expliquer (Binaire : Bernoulli, Comptage : Poisson, continue : normale ou gamma).
2. Fonction de lien. Ce choix est plus délicat. La fonction de lien dite "canonique" $g(u) = (b')^{-1}(u)$ est souvent privilégiée (notamment pour des raisons d'écriture de modèles et de simplicité d'écriture)

Propriété

Les fonctions de lien des modèles logistique et linéaire sont canoniques.

Choix de la loi exponentielle et de la fonction de lien

1. Loi exponentielle. Ce choix est généralement guidé par la nature de la variable à expliquer (Binaire : Bernoulli, Comptage : Poisson, continue : normale ou gamma).
2. Fonction de lien. Ce choix est plus délicat. La fonction de lien dite "canonique" $g(u) = (b')^{-1}(u)$ est souvent privilégiée (notamment pour des raisons d'écriture de modèles et de simplicité d'écriture)

Propriété

Les fonctions de lien des modèles logistique et linéaire sont canoniques.

Choix de la loi exponentielle et de la fonction de lien

1. Loi exponentielle. Ce choix est généralement guidé par la nature de la variable à expliquer (Binaire : Bernoulli, Comptage : Poisson, continue : normale ou gamma).

2. Fonction de lien. Ce choix est plus délicat. La fonction de lien dite "canonique" $g(u) = (b')^{-1}(u)$ est souvent privilégiée (notamment pour des raisons d'écriture de modèles et de simplicité d'écriture)

Propriété

Les fonctions de lien des modèles logistique et linéaire sont canoniques.

Fonctions de lien usuelles

Nom du lien	Fonction de lien
identité	$g(u) = u$
log	$g(u) = \log(u)$
cloglog	$g(u) = \log(-\log(1-u))$
logit	$g(u) = log(u/(1-u))$
probit	$g(u) = \Phi^{-1}(u)$
réciproque	$g(u) = -1/u$
puissance	$g(u) = u^{\gamma}$, $\gamma \neq 0$

GLM sur R

- Il faut bien entendu spécifier à la fonction glm les 3 composantes d'un modèle glm :

glm(formula=....,family=....(link=....))

1. formula : spécifie la composante déterministe $Y = X_1 + X_2$, $Y = X_1 + X_2 + X_1 : X_2$ (prendre en compte l'interaction entre X_1 et X_2.
2. family : spécifie composante aléatoire (gaussian pour le modèle linéaire gaussien, binomial lorsque la variable à expliquer est binaire...)
3. link : spécifie la fonction de lien (logit pour logistique, probit pour probit...)

GLM sur R

- Il faut bien entendu spécifier à la fonction glm les 3 composantes d'un modèle glm :

glm(formula=....,family=....(link=....))

1. formula : spécifie la composante déterministe $Y = X_1 + X_2$, $Y = X_1 + X_2 + X_1 : X_2$ (prendre en compte l'interaction entre X_1 et X_2.
2. family : spécifie composante aléatoire (gaussian pour le modèle linéaire gaussien, binomial lorsque la variable à expliquer est binaire...)
3. link : spécifie la fonction de lien (logit pour logistique, probit pour probit...)

GLM sur R

- Il faut bien entendu spécifier à la fonction glm les 3 composantes d'un modèle glm :

glm(formula=....,family=....(link=....))

1. formula : spécifie la composante déterministe $Y = X_1 + X_2$, $Y = X_1 + X_2 + X_1 : X_2$ (prendre en compte l'interaction entre X_1 et X_2.
2. family : spécifie composante aléatoire (gaussian pour le modèle linéaire gaussien, binomial lorsque la variable à expliquer est binaire...)
3. link : spécifie la fonction de lien (logit pour logistique, probit pour probit...)

GLM sur R

- Il faut bien entendu spécifier à la fonction glm les 3 composantes d'un modèle glm :

glm(formula=....,family=....(link=....))

1. formula : spécifie la composante déterministe $Y = X_1 + X_2$, $Y = X_1 + X_2 + X_1 : X_2$ (prendre en compte l'interaction entre X_1 et X_2.
2. family : spécifie composante aléatoire (gaussian pour le modèle linéaire gaussien, binomial lorsque la variable à expliquer est binaire...)
3. link : spécifie la fonction de lien (logit pour logistique, probit pour probit...)

Exemple

- On cherche à expliquer une variable binaire Y par deux variables continues X_1 et X_2 :

```
> Y <- rbinom(50,1,0.6)
> X1 <- runif(50)
> X2 <- rnorm(50)
```

- On ajuste les modèles

```
> glm(Y~X1+X2,family=binomial)
Coefficients:
(Intercept)           X1           X2
    -0.2849       1.8610      -0.0804

> glm(Y~X1+X2+X1:X2,family=binomial)
Coefficients:
(Intercept)           X1           X2        X1:X2
    -0.3395       2.1175      -0.4568       1.0346

> glm(Y~X1+X2,family=binomial(link = "probit"))
Coefficients:
(Intercept)           X1           X2
   -0.17038      1.11986     -0.04864
```

Le problème

- On cherche à quantifier l'influence d'un traitement sur l'évolution du nombre de polypes au colon. On dispose des données suivantes :

```
   number   treat age
1      63 placebo  20
2       2    drug  16
3      28 placebo  18
4      17    drug  22
5      61 placebo  13
...
```

où

  - `number` : nombre de polypes après 12 mois de traitement.
  - `treat` : drug si le traitement a été administré, placebo sinon.
  - age : age de l'individu.

Le problème est d'expliquer la variable `number` par les deux autres variables à l'aide d'un GLM.

Le problème

- On cherche à quantifier l'influence d'un traitement sur l'évolution du nombre de polypes au colon. On dispose des données suivantes :

```
   number   treat age
1      63 placebo  20
2       2    drug  16
3      28 placebo  18
4      17    drug  22
5      61 placebo  13
...
```

où

  - `number` : nombre de polypes après 12 mois de traitement.
  - `treat` : drug si le traitement a été administré, placebo sinon.
  - age : age de l'individu.

Le problème est d'expliquer la variable `number` par les deux autres variables à l'aide d'un GLM.

Notations

On note

- Y_i la variable aléatoire représentant le nombre de polypes du $i$ème patient après les 12 mois de traitement.
- x_{i1} la variable `treat` pour le $i$ème individu et x_{i2} l'age du $i$ème individu.

GLM

1. La variable Y_i étant une variable de comptage, on choisit comme densité de Y_i la densité (par rapport à la mesure de comptage) de la loi de Poisson de paramètre λ_i :

$$f_{\alpha_i}(y_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!} = \exp[y_i \log(\lambda_i) - \exp(\log(\lambda_i)) - \log(y_i!)].$$

2. La fonction de lien canonique est donc donnée par :

$$g(u) = \log(u).$$

Notations

On note

- Y_i la variable aléatoire représentant le nombre de polypes du $i$ème patient après les 12 mois de traitement.
- x_{i1} la variable `treat` pour le $i$ème individu et x_{i2} l'age du $i$ème individu.

GLM

1. La variable Y_i étant une variable de comptage, on choisit comme densité de Y_i la densité (par rapport à la mesure de comptage) de la loi de Poisson de paramètre λ_i :

$$f_{\alpha_i}(y_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!} = \exp[y_i \log(\lambda_i) - \exp(\log(\lambda_i)) - \log(y_i!)].$$

2. La fonction de lien canonique est donc donnée par :

$$g(u) = \log(u).$$

Modèle de Poisson ou régression log-linéaire

Définition

Le modèle de Poisson modélise la loi de Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que

$$\log(\lambda(x_i)) = x_i'\beta.$$

- L'ajustement sur R s'effectue toujours à l'aide de la fonction glm :

```
> glm(number~treat+age,data=polyps,family=poisson)

Coefficients:
(Intercept)    treatdrug          age
    4.52902     -1.35908     -0.03883
```

- **Prédiction** : pour un individu de 23 ans, ayant reçu le traitement on pourra estimer le nombre de polypes à 12 mois par

$$\exp(4.52902 - 1.35908 - 0.03883 * 23) = 9.745932.$$

Modèle de Poisson ou régression log-linéaire

Définition

Le modèle de Poisson modélise la loi de Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que

$$\log(\lambda(x_i)) = x_i'\beta.$$

- L'ajustement sur R s'effectue toujours à l'aide de la fonction glm :

```
> glm(number~treat+age,data=polyps,family=poisson)

Coefficients:
(Intercept)    treatdrug          age
    4.52902     -1.35908     -0.03883
```

- **Prédiction** : pour un individu de 23 ans, ayant reçu le traitement on pourra estimer le nombre de polypes à 12 mois par

$$\exp(4.52902 - 1.35908 - 0.03883 * 23) = 9.745932.$$

Modèle de Poisson ou régression log-linéaire

Définition

Le modèle de Poisson modélise la loi de Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que

$$\log(\lambda(x_i)) = x_i'\beta.$$

- L'ajustement sur R s'effectue toujours à l'aide de la fonction glm :

```
> glm(number~treat+age,data=polyps,family=poisson)

Coefficients:
(Intercept)    treatdrug          age
    4.52902     -1.35908     -0.03883
```

- **Prédiction** : pour un individu de 23 ans, ayant reçu le traitement on pourra estimer le nombre de polypes à 12 mois par

$$\exp(4.52902 - 1.35908 - 0.03883 * 23) = 9.745932.$$

Modèle de Poisson ou régression log-linéaire

Définition

Le modèle de Poisson modélise la loi de Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que

$$\log(\lambda(x_i)) = x_i'\beta.$$

- L'ajustement sur R s'effectue toujours à l'aide de la fonction glm :

```
> glm(number~treat+age,data=polyps,family=poisson)

Coefficients:
(Intercept)    treatdrug          age
    4.52902     -1.35908     -0.03883
```

- **Prédiction** : pour un individu de 23 ans, ayant reçu le traitement on pourra estimer le nombre de polypes à 12 mois par

$$\exp(4.52902 - 1.35908 - 0.03883 * 23) = 9.745932.$$

Deuxième partie II

Analyse du modèle de régression logistique

1. Le modèle
   - Présentation
   - Identifiabilité et la matrice de design
   - Interprétation des coefficients

2. Estimation des paramètres
   - La vraisemblance
   - Existence et unicité de l'emv
   - L'algorithme IRLS
   - Comportement asymptotique de l'emv

3. Bibliographie

Notations

- On cherche à expliquer une variable Y binaire par p variables explicatives $X_1, \ldots, X_p$.
- On dispose de n observations $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in \mathbb{R}^p$, $y_i \in \{0, 1\}$.
- On note $\mathbb{X}$ la matrice $n \times p$ contenant les observations des variables explicatives :

$$\mathbb{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Le modèle de régression logistique

- On suppose que les variables explicatives sont déterministes.

Modèle logistique, [Hosmer and Lemeshow, 2000]

Les observations y_i sont des réalisations de variables aléatoires Y_i indépendantes de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1-p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- **Remarque :** la variable X_1 peut correspondre à la constante du modèle. Dans ce cas, $x_{i1} = 1, i = 1, \ldots, n$.

Le modèle de régression logistique

- On suppose que les variables explicatives sont déterministes.

Modèle logistique, [Hosmer and Lemeshow, 2000]

Les observations y_i sont des réalisations de variables aléatoires Y_i indépendantes de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1-p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- **Remarque :** la variable X_1 peut correspondre à la constante du modèle. Dans ce cas, $x_{i1} = 1, i = 1, \ldots, n$.

Le modèle de régression logistique

- On suppose que les variables explicatives sont déterministes.

Modèle logistique, [Hosmer and Lemeshow, 2000]

Les observations y_i sont des réalisations de variables aléatoires Y_i indépendantes de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1-p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- **Remarque :** la variable X_1 peut correspondre à la constante du modèle. Dans ce cas, $x_{i1} = 1, i = 1, \ldots, n$.

Autre présentation du modèle logistique

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

Autre présentation du modèle logistique

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

Autre présentation du modèle logistique

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

alors le modèle est le modèle logistique.

- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

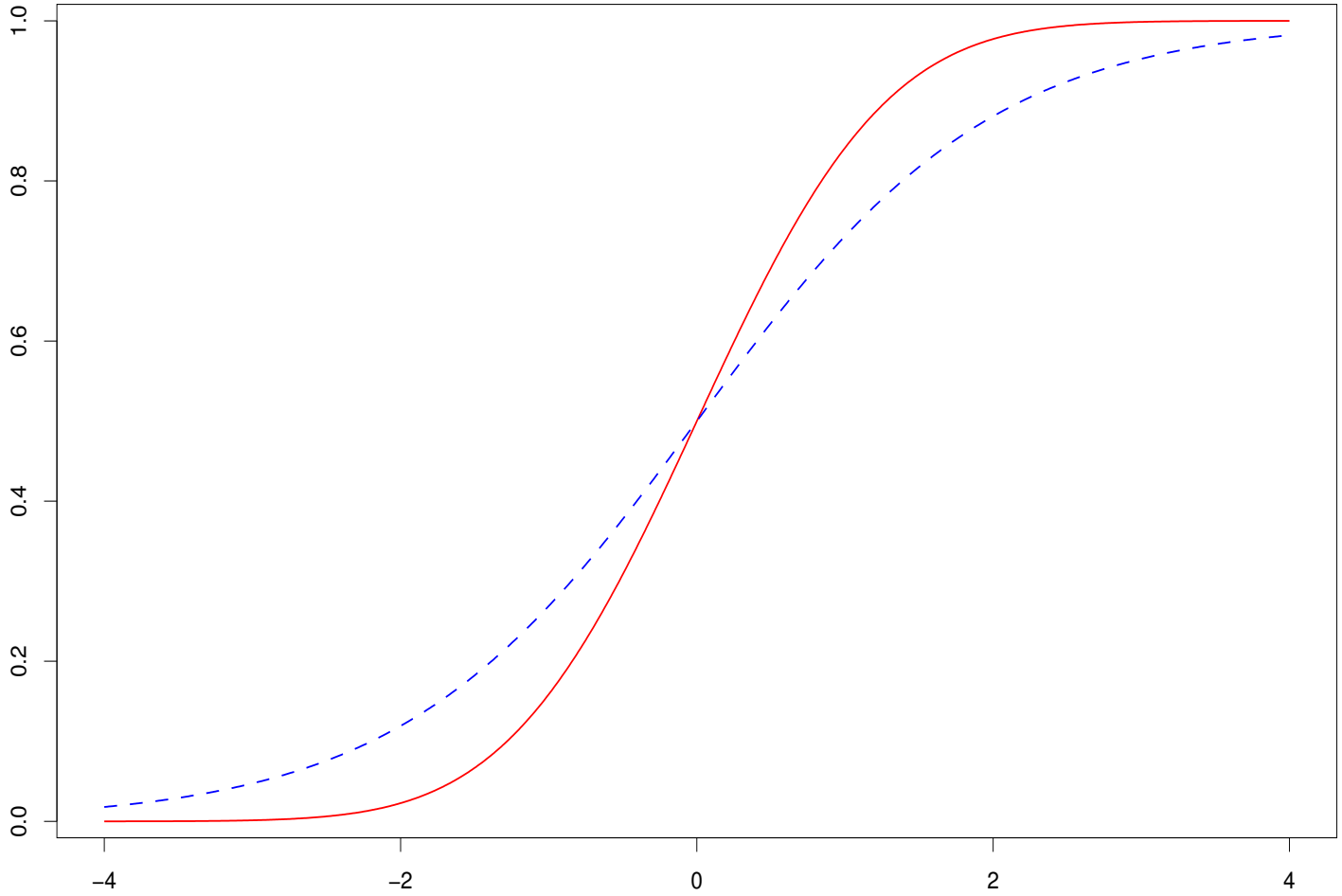


Figure – Fonctions de répartition pour le modèle logistique (bleu) et probit (rouge).

2 types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

2 types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

2 types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

2 types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

2 types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

Données répétées

On note toujours n le nombre d'observations. On note

- $x_1, \ldots, x_T$ les différentes valeurs des variables explicatives observées.
- $n_1, \ldots, n_T$ tel que $n_t =$ nombre de fois où x_t a été observé : $\sum_{t=1}^{T} n_t = n$.
- $y_1, \ldots, y_T$ les nombres de succès observés au point x_t.

Modèle logistique pour données répétées

Si on suppose que y_t est une réalisation d'une variable aléatoire Y_t alors cette fois la loi de Y_t n'est plus une Bernoulli mais une binomiale de paramètre $(n_t, p_\beta(x_t))$ tel que

$$\text{logit}\, p_\beta(x_t) = \beta_1 x_{t1} + \ldots + \beta_p x_{tp} = x_t' \beta.$$

- **Remarque** : le cas données individuelles est un cas particulier de données répétées (il suffit de poser $T = n$).

Données répétées

On note toujours n le nombre d'observations. On note

- $x_1, \ldots, x_T$ les différentes valeurs des variables explicatives observées.
- $n_1, \ldots, n_T$ tel que $n_t =$ nombre de fois où x_t a été observé : $\sum_{t=1}^{T} n_t = n$.
- $y_1, \ldots, y_T$ les nombres de succés observés au point x_t.

Modèle logistique pour données répétées

Si on suppose que y_t est une réalisation d'une variable aléatoire Y_t alors cette fois la loi de Y_t n'est plus une Bernoulli mais une binomiale de paramètre $(n_t, p_\beta(x_t))$ tel que

$$\text{logit}\, p_\beta(x_t) = \beta_1 x_{t1} + \ldots + \beta_p x_{tp} = x_t' \beta.$$

- **Remarque** : le cas données individuelles est un cas particulier de données répétées (il suffit de poser $T = n$).

Données répétées

On note toujours n le nombre d'observations. On note

- $x_1, \ldots, x_T$ les différentes valeurs des variables explicatives observées.
- $n_1, \ldots, n_T$ tel que $n_t =$ nombre de fois où x_t a été observé : $\sum_{t=1}^{T} n_t = n$.
- $y_1, \ldots, y_T$ les nombres de succés observés au point x_t.

Modèle logistique pour données répétées

Si on suppose que y_t est une réalisation d'une variable aléatoire Y_t alors cette fois la loi de Y_t n'est plus une Bernoulli mais une binomiale de paramètre $(n_t, p_\beta(x_t))$ tel que

$$\text{logit}\, p_\beta(x_t) = \beta_1 x_{t1} + \ldots + \beta_p x_{tp} = x_t' \beta.$$

- **Remarque** : le cas données individuelles est un cas particulier de données répétées (il suffit de poser $T = n$).

Données répétées

On note toujours n le nombre d'observations. On note

- $x_1, \ldots, x_T$ les différentes valeurs des variables explicatives observées.
- $n_1, \ldots, n_T$ tel que $n_t =$ nombre de fois où x_t a été observé : $\sum_{t=1}^{T} n_t = n$.
- $y_1, \ldots, y_T$ les nombres de succés observés au point x_t.

Modèle logistique pour données répétées

Si on suppose que y_t est une réalisation d'une variable aléatoire Y_t alors cette fois la loi de Y_t n'est plus une Bernoulli mais une binomiale de paramètre $(n_t, p_\beta(x_t))$ tel que

$$\text{logit}\, p_\beta(x_t) = \beta_1 x_{t1} + \ldots + \beta_p x_{tp} = x_t' \beta.$$

- **Remarque** : le cas données individuelles est un cas particulier de données répétées (il suffit de poser $T = n$).

Données répétées

On note toujours n le nombre d'observations. On note

- $x_1, \ldots, x_T$ les différentes valeurs des variables explicatives observées.
- $n_1, \ldots, n_T$ tel que $n_t =$ nombre de fois où x_t a été observé : $\sum_{t=1}^{T} n_t = n$.
- $y_1, \ldots, y_T$ les nombres de succés observés au point x_t.

Modèle logistique pour données répétées

Si on suppose que y_t est une réalisation d'une variable aléatoire Y_t alors cette fois la loi de Y_t n'est plus une Bernoulli mais une binomiale de paramètre $(n_t, p_\beta(x_t))$ tel que

$$\text{logit}\, p_\beta(x_t) = \beta_1 x_{t1} + \ldots + \beta_p x_{tp} = x_t' \beta.$$

- **Remarque** : le cas données individuelles est un cas particulier de données répétées (il suffit de poser $T = n$).

Identifiabilité

- **Rappels :** la régression logistique modélise la loi de Y_i par une Bernoulli de paramètre $p_\beta(x_i)$.
- Par définition, le modèle est dit identifiable si $\beta \mapsto \mathbf{P}_{(Y_1,\ldots,Y_n)}$ est injective.
- Le modèle logistique est donc identifiable si pour tout $\beta \neq \tilde{\beta}$ il existe $i \in \{1,\ldots,n\}$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$.

Propriété

Si $n > p$ alors le modèle est identifiable si et seulement si $\text{rang}(\mathbb{X}) = p$.

Identifiabilité

- **Rappels :** la régression logistique modélise la loi de Y_i par une Bernoulli de paramètre $p_\beta(x_i)$.
- Par définition, le modèle est dit identifiable si $\beta \mapsto \mathbf{P}_{(Y_1,\ldots,Y_n)}$ est injective.
- Le modèle logistique est donc identifiable si pour tout $\beta \neq \tilde{\beta}$ il existe $i \in \{1,\ldots,n\}$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$.

Propriété

Si $n > p$ alors le modèle est identifiable si et seulement si $\text{rang}(\mathbb{X}) = p$.

Identifiabilité

- **Rappels :** la régression logistique modélise la loi de Y_i par une Bernoulli de paramètre $p_\beta(x_i)$.
- Par définition, le modèle est dit identifiable si $\beta \mapsto \mathbf{P}_{(Y_1,\ldots,Y_n)}$ est injective.
- Le modèle logistique est donc identifiable si pour tout $\beta \neq \tilde{\beta}$ il existe $i \in \{1,\ldots,n\}$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$.

Propriété

Si $n > p$ alors le modèle est identifiable si et seulement si $\text{rang}(\mathbb{X}) = p$.

Identifiabilité

- **Rappels :** la régression logistique modélise la loi de Y_i par une Bernoulli de paramètre $p_\beta(x_i)$.
- Par définition, le modèle est dit identifiable si $\beta \mapsto \mathbf{P}_{(Y_1,\ldots,Y_n)}$ est injective.
- Le modèle logistique est donc identifiable si pour tout $\beta \neq \tilde{\beta}$ il existe $i \in \{1,\ldots,n\}$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$.

Propriété

Si $n > p$ alors le modèle est identifiable si et seulement si $\text{rang}(\mathbb{X}) = p$.

Rappels

- La matrice de design $\mathbb{X}$ contient "les observations des variables explicatives" :

$$\mathbb{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

- Elle joue un role important pour :
  1. l'identifiabilité du modèle ;
  2. l'estimation des paramètres du modèle ;
  3. le comportement asymptotique des estimateurs du modèle.

Rappels

- La matrice de design $\mathbb{X}$ contient "les observations des variables explicatives" :

$$\mathbb{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

- Elle joue un role important pour :
  1. l'identifiabilité du modèle ;
  2. l'estimation des paramètres du modèle ;
  3. le comportement asymptotique des estimateurs du modèle.

Un exemple

- Un chef d'entreprise souhaite vérifier la qualité d'un type de machines en fonction de l'âge et de la marque des moteurs. Il dispose
  1. d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
  2. d'une variable quantitative `age` repésentant l'âge du moteur ;
  3. d'une variable qualitative a 3 modalités `marque` représentant la marque du moteur,

- et de $n = 33$ observations :

```
> panne
   etat age marque
1     0   4      A
2     0   2      C
3     0   3      C
4     0   9      B
5     0   7      B
```

Un exemple

- Un chef d'entreprise souhaite vérifier la qualité d'un type de machines en fonction de l'âge et de la marque des moteurs. Il dispose
  1. d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
  2. d'une variable quantitative age repésentant l'âge du moteur ;
  3. d'une variable qualitative a 3 modalités `marque` représentant la marque du moteur,

- et de $n = 33$ observations :

```
> panne
   etat age marque
1     0   4      A
2     0   2      C
3     0   3      C
4     0   9      B
5     0   7      B
```

Variable quantitative

- C'est le cas le plus simple, un seul coefficient est dans le modèle par variable explicative $\Longrightarrow$ une variable quantitative est représentée par une seule colonne dans la matrice de design.

Exemple : pannes de machines

- On considère les modèles logistiques permettant d'expliquer `panne` par age et par age et la constante :

$$\text{logit}\, p_\beta(x_i) = \beta x_i \quad \text{et} \quad \text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Les matrice de design associées à ces deux modèles sont

$$\mathbb{X} = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 9 \\ \vdots \end{pmatrix} \quad \text{et} \quad \mathbb{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 3 \\ 1 & 9 \\ \vdots & \vdots \end{pmatrix}.$$

Variable quantitative

- C'est le cas le plus simple, un seul coefficient est dans le modèle par variable explicative $\Longrightarrow$ une variable quantitative est représentée par une seule colonne dans la matrice de design.

Exemple : pannes de machines

- On considère les modèles logistiques permettant d'expliquer `panne` par age et par age et la constante :

$$\text{logit}\, p_\beta(x_i) = \beta x_i \quad \text{et} \quad \text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Les matrice de design associées à ces deux modèles sont

$$\mathbb{X} = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 9 \\ \vdots \end{pmatrix} \quad \text{et} \quad \mathbb{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 3 \\ 1 & 9 \\ \vdots & \vdots \end{pmatrix}.$$

Variable quantitative

- C'est le cas le plus simple, un seul coefficient est dans le modèle par variable explicative $\Longrightarrow$ une variable quantitative est représentée par une seule colonne dans la matrice de design.

Exemple : pannes de machines

- On considère les modèles logistiques permettant d'expliquer `panne` par age et par age et la constante :

$$\text{logit}\, p_\beta(x_i) = \beta x_i \quad \text{et} \quad \text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Les matrice de design associées à ces deux modèles sont

$$\mathbb{X} = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 9 \\ \vdots \end{pmatrix} \quad \text{et} \quad \mathbb{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 3 \\ 1 & 9 \\ \vdots & \vdots \end{pmatrix}.$$

Variable qualitative

- Considérons le modèle logistique avec pour variable explicative `marque` pour l'exemple des **pannes de machines**. Une écriture naturelle est :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_i) + \beta_2 \mathbf{1}_B(x_i) + \beta_3 \mathbf{1}_C(x_i).$$

- Ce modèle n'est clairement pas identifiable.
- En effet, la matrice de design associée à ce modèle

$$\begin{bmatrix} A \\ C \\ C \\ B \\ B \\ \vdots \end{bmatrix} \Longrightarrow \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

n'est clairement pas de plein rang.

Variable qualitative

- Considérons le modèle logistique avec pour variable explicative `marque` pour l'exemple des **pannes de machines**. Une écriture naturelle est :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_i) + \beta_2 \mathbf{1}_B(x_i) + \beta_3 \mathbf{1}_C(x_i).$$

- Ce modèle n'est clairement pas identifiable.
- En effet, la matrice de design associée à ce modèle

$$\begin{bmatrix} A \\ C \\ C \\ B \\ B \\ \vdots \end{bmatrix} \Longrightarrow \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

n'est clairement pas de plein rang.

Variable qualitative

- Considérons le modèle logistique avec pour variable explicative `marque` pour l'exemple des **pannes de machines**. Une écriture naturelle est :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_i) + \beta_2 \mathbf{1}_B(x_i) + \beta_3 \mathbf{1}_C(x_i).$$

- Ce modèle n'est clairement pas identifiable.
- En effet, la matrice de design associée à ce modèle

$$\begin{bmatrix} A \\ C \\ C \\ B \\ B \\ \vdots \end{bmatrix} \Longrightarrow \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

n'est clairement pas de plein rang.

Contraintes d'identifiabilité

- Le modèle précédent est surparamétré. Il est nécessaire de définir des contraintes d'identifiabilité.
- Les contraintes les plus utilisées sont $\beta_1 = 0$ (choix une modalité de référence) ou $\beta_1 + \beta_2 + \beta_3 = 0$.
- Le choix de la contrainte n'est pas forcément spécifié dans les sorties logiciels. Il est capital d'aller voir l'aide des fonctions pour connaitre les contraintes d'identifiabilité.

Remarque

Mathématiquement, le choix de la contrainte n'a pas une grande importance. Il doit en revanche être pris en compte pour pouvoir interpréter correctement les paramètres du modèle.

Contraintes d'identifiabilité

- Le modèle précédent est surparamétré. Il est nécessaire de définir des contraintes d'identifiabilité.
- Les contraintes les plus utilisées sont $\beta_1 = 0$ (choix une modalité de référence) ou $\beta_1 + \beta_2 + \beta_3 = 0$.
- Le choix de la contrainte n'est pas forcément spécifié dans les sorties logiciels. Il est capital d'aller voir l'aide des fonctions pour connaitre les contraintes d'identifiabilité.

Remarque

Mathématiquement, le choix de la contrainte n'a pas une grande importance. Il doit en revanche être pris en compte pour pouvoir interpréter correctement les paramètres du modèle.

Contraintes d'identifiabilité

- Le modèle précédent est surparamétré. Il est nécessaire de définir des contraintes d'identifiabilité.
- Les contraintes les plus utilisées sont $\beta_1 = 0$ (choix une modalité de référence) ou $\beta_1 + \beta_2 + \beta_3 = 0$.
- Le choix de la contrainte n'est pas forcément spécifié dans les sorties logiciels. Il est capital d'aller voir l'aide des fonctions pour connaitre les contraintes d'identifiabilité.

Remarque

Mathématiquement, le choix de la contrainte n'a pas une grande importance. Il doit en revanche être pris en compte pour pouvoir interpréter correctement les paramètres du modèle.

Contraintes d'identifiabilité

- Le modèle précédent est surparamétré. Il est nécessaire de définir des contraintes d'identifiabilité.
- Les contraintes les plus utilisées sont $\beta_1 = 0$ (choix une modalité de référence) ou $\beta_1 + \beta_2 + \beta_3 = 0$.
- Le choix de la contrainte n'est pas forcément spécifié dans les sorties logiciels. Il est capital d'aller voir l'aide des fonctions pour connaitre les contraintes d'identifiabilité.

Remarque

Mathématiquement, le choix de la contrainte n'a pas une grande importance. Il doit en revanche être pris en compte pour pouvoir interpréter correctement les paramètres du modèle.

Contrainte d'identifiabilité sous R

- Par défaut, R choisit comme modalité de référence la première modalité de la variable qualitative :

```
> glm(etat~marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC
     0.5596      -0.4261      -1.4759
```

- On peut modifier le choix de la modalité de référence

```
> glm(etat~C(marque,base=2),data=panne,family=binomial)

Coefficients:
         (Intercept)  C(marque, base = 2)1  C(marque, base = 2)3
              0.1335                0.4261               -1.0498
```

Contrainte d'identifiabilité sous R

- Par défaut, R choisit comme modalité de référence la première modalité de la variable qualitative :

```
> glm(etat~marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC
     0.5596      -0.4261      -1.4759
```

- On peut modifier le choix de la modalité de référence

```
> glm(etat~C(marque,base=2),data=panne,family=binomial)

Coefficients:
         (Intercept)  C(marque, base = 2)1  C(marque, base = 2)3
              0.1335                0.4261               -1.0498
```

Intéractions

Définition

Deux variables explicatives interagissent si l'effet de l'une de ces variables sur la variable à expliquer est différent selon les modalités de l'autre.

- On considère le modèle logistique permettant d'expliquer Y (`etat`) par X_1 (`marque`) et X_2 (age) :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_{i1}) + \beta_2 \mathbf{1}_B(x_{i1}) + \beta_3 \mathbf{1}_C(x_{i1}) + \beta_4 x_{i2},$$

  muni de la contrainte $\beta_1 = 0$.
- Ce modèle stipule que l'age de la machine agit linéairement sur $\text{logit}\, p_\beta(x_i)$ et que le coefficient de linéarité est le même pour toutes les marques.

- Il est bien entendu possible d'envisager que l'effet de l'age sur l'état de la machine ne soit pas exactement le même pour toutes les marques :

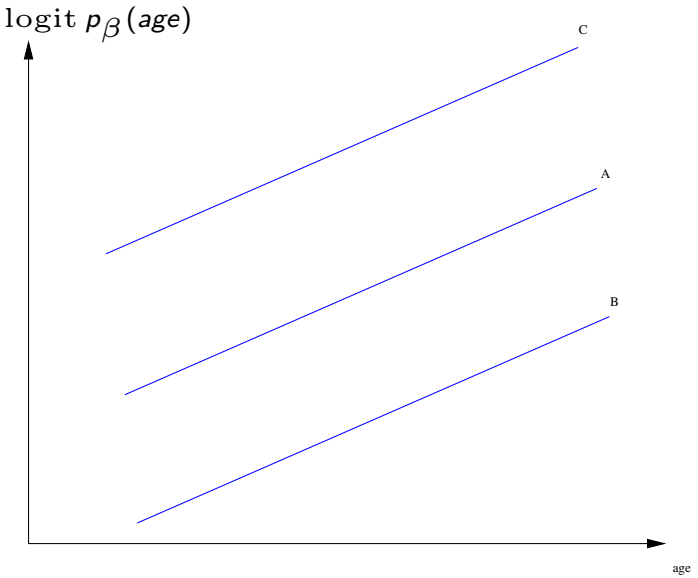


Table – Modèle additif.

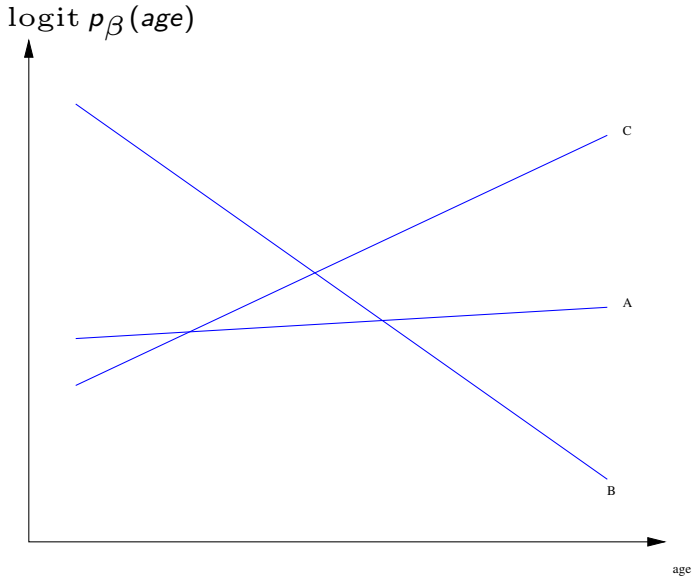


Table – Modèle avec interaction.

- La figure de droite correspond à un modèle du genre

$$\text{logit}(p_\beta(x_i)) = \begin{cases} \beta_{01} + \beta_{11} x_{i2} & \text{si } x_{i1} = A \\ \beta_{02} + \beta_{12} x_{i2} & \text{si } x_{i1} = B \\ \beta_{03} + \beta_{13} x_{i2} & \text{si } x_{i1} = C. \end{cases} \tag{2}$$

- Un tel modèle prend en compte l'interaction `marque:age` :

```
> glm(etat~marque+age+age:marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC          age  marqueB:age  marqueC:age
    0.23512      0.19862     -2.43145      0.05641     -0.11188      0.21587
```

- Le modèle ajusté ici n'est pas exactement celui défini par (2). Il est paramétré différemment :

$$\begin{aligned} \text{logit}(p_\beta(x_i)) = & \gamma_0 + \gamma_1 \mathbf{1}_A(x_{i1}) + \gamma_2 \mathbf{1}_B(x_{i1}) + \gamma_3 \mathbf{1}_C(x_{i1}) + \gamma_4 x_{i2} \\ & + \gamma_5 x_{i2} \mathbf{1}_A(x_{i1}) + \gamma_6 x_{i2} \mathbf{1}_B(x_{i1}) + \gamma_7 x_{i2} \mathbf{1}_C(x_{i1}) \end{aligned}$$

  muni des contraintes $\gamma_1 = 0$ et $\gamma_5 = 0$.

- La figure de droite correspond à un modèle du genre

$$\text{logit}(p_\beta(x_i)) = \begin{cases} \beta_{01} + \beta_{11} x_{i2} & \text{si } x_{i1} = A \\ \beta_{02} + \beta_{12} x_{i2} & \text{si } x_{i1} = B \\ \beta_{03} + \beta_{13} x_{i2} & \text{si } x_{i1} = C. \end{cases} \tag{2}$$

- Un tel modèle prend en compte l'interaction `marque:age` :

```
> glm(etat~marque+age+age:marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC          age  marqueB:age  marqueC:age
    0.23512      0.19862     -2.43145      0.05641     -0.11188      0.21587
```

- Le modèle ajusté ici n'est pas exactement celui défini par (2). Il est paramétré différemment :

$$\begin{aligned} \text{logit}(p_\beta(x_i)) = & \gamma_0 + \gamma_1 \mathbf{1}_A(x_{i1}) + \gamma_2 \mathbf{1}_B(x_{i1}) + \gamma_3 \mathbf{1}_C(x_{i1}) + \gamma_4 x_{i2} \\ & + \gamma_5 x_{i2} \mathbf{1}_A(x_{i1}) + \gamma_6 x_{i2} \mathbf{1}_B(x_{i1}) + \gamma_7 x_{i2} \mathbf{1}_C(x_{i1}) \end{aligned}$$

  muni des contraintes $\gamma_1 = 0$ et $\gamma_5 = 0$.

- La figure de droite correspond à un modèle du genre

$$\text{logit}(p_\beta(x_i)) = \begin{cases} \beta_{01} + \beta_{11} x_{i2} & \text{si } x_{i1} = A \\ \beta_{02} + \beta_{12} x_{i2} & \text{si } x_{i1} = B \\ \beta_{03} + \beta_{13} x_{i2} & \text{si } x_{i1} = C. \end{cases} \tag{2}$$

- Un tel modèle prend en compte l'interaction `marque:age` :

```
> glm(etat~marque+age+age:marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC          age  marqueB:age  marqueC:age
    0.23512      0.19862     -2.43145      0.05641     -0.11188      0.21587
```

- Le modèle ajusté ici n'est pas exactement celui défini par (2). Il est paramétré différemment :

$$\begin{aligned} \text{logit}(p_\beta(x_i)) = &\gamma_0 + \gamma_1 \mathbf{1}_A(x_{i1}) + \gamma_2 \mathbf{1}_B(x_{i1}) + \gamma_3 \mathbf{1}_C(x_{i1}) + \gamma_4 x_{i2} \\ &+ \gamma_5 x_{i2} \mathbf{1}_A(x_{i1}) + \gamma_6 x_{i2} \mathbf{1}_B(x_{i1}) + \gamma_7 x_{i2} \mathbf{1}_C(x_{i1}) \end{aligned}$$

  muni des contraintes $\gamma_1 = 0$ et $\gamma_5 = 0$.

- Il est facile de voir que les coefficients β_{jk} se déduisent des cofficients γ_j. Par exemple

$$\beta_{01} = \gamma_0 + \gamma_1, \quad \beta_{12} = \gamma_4 + \gamma_6, \quad \dots$$

- On peut bien évidemment ajuster directement le modèle (2) :

```
> glm(etat~marque-1+age:marque,data=panne,family=binomial)

Coefficients:
    marqueA      marqueB      marqueC  marqueA:age  marqueB:age  marqueC:age
    0.23512      0.43375     -2.19633      0.05641     -0.05547      0.27228
```

- L'interaction présentée ci-dessus met en jeu une variable quantitative (age) avec une variable qualitative (marque). Il est bien entendu possible d'inclure des interactions entre variables qualitatives ou (plus rare) quantitatives.

- Il est facile de voir que les coefficients β_{jk} se déduisent des cofficients γ_j. Par exemple

$$\beta_{01} = \gamma_0 + \gamma_1, \quad \beta_{12} = \gamma_4 + \gamma_6, \quad \ldots$$

- On peut bien évidemment ajuster directement le modèle (2) :

```
> glm(etat~marque-1+age:marque,data=panne,family=binomial)

Coefficients:
    marqueA      marqueB      marqueC  marqueA:age  marqueB:age  marqueC:age
    0.23512      0.43375     -2.19633      0.05641     -0.05547      0.27228
```

- L'interaction présentée ci-dessus met en jeu une variable quantitative (age) avec une variable qualitative (marque). Il est bien entendu possible d'inclure des interactions entre variables qualitatives ou (plus rare) quantitatives.

- Il est facile de voir que les coefficients β_{jk} se déduisent des cofficients γ_j. Par exemple

$$\beta_{01} = \gamma_0 + \gamma_1, \quad \beta_{12} = \gamma_4 + \gamma_6, \quad \ldots$$

- On peut bien évidemment ajuster directement le modèle (2) :

```
> glm(etat~marque-1+age:marque,data=panne,family=binomial)

Coefficients:
    marqueA      marqueB      marqueC  marqueA:age  marqueB:age  marqueC:age
    0.23512      0.43375     -2.19633      0.05641     -0.05547      0.27228
```

- L'interaction présentée ci-dessus met en jeu une variable quantitative (age) avec une variable qualitative (marque). Il est bien entendu possible d'inclure des interactions entre variables qualitatives ou (plus rare) quantitatives.

Dimension d'un modèle

Définition

La dimension d'un modèle (logistique) est égale au nombre de paramètres identifiables du modèle. Elle correspond au nombre de colonnes de la matrice de design $\mathbb{X}$.

On calcule la dimension du modèle en sommant les contributions de chaque variables du modèle :

- si la constante est présente, elle a une contribution de 1 :
- une variable quantitative (non discrétisée) a une contribution de 1 ;
- une variable qualitative à K modalités a une contribution de $K - 1$;
- la contribution d'une interaction s'obtient en faisant le produit des contributions des variables qui interagissent.

Dimension d'un modèle

Définition

La dimension d'un modèle (logistique) est égale au nombre de paramètres identifiables du modèle. Elle correspond au nombre de colonnes de la matrice de design $\mathbb{X}$.

On calcule la dimension du modèle en sommant les contributions de chaque variables du modèle :

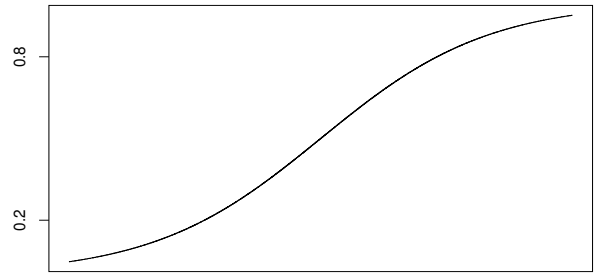
- si la constante est présente, elle a une contribution de 1 :
- une variable quantitative (non discrétisée) a une contribution de 1 ;
- une variable qualitative à K modalités a une contribution de $K - 1$;
- la contribution d'une interaction s'obtient en faisant le produit des contributions des variables qui interagissent.

- On considère l'allure de la courbe représentative de

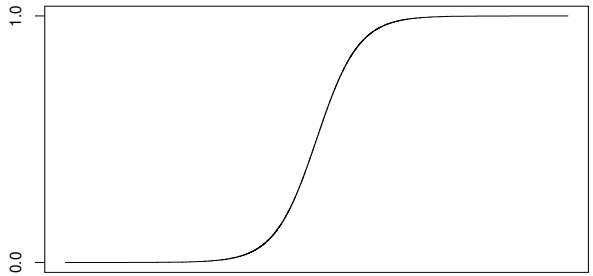
$$x \mapsto p_\beta(x) = \frac{\exp(x\beta)}{1+\exp(x\beta)}$$

pour $\beta = 0, 0.5, 2, 10$

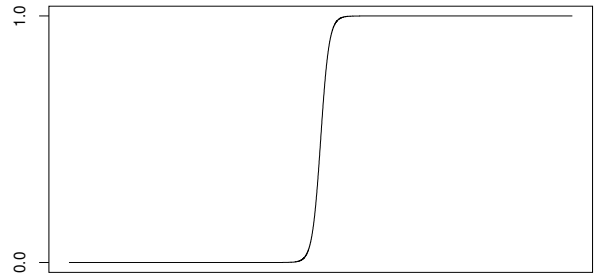
beta0



beta0.5



beta2



beta10

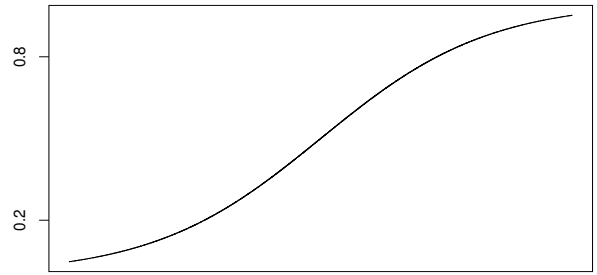
Lorsque β augmente, $p_\beta(x)$ est souvent proche de 0 ou 1.

- On considère l'allure de la courbe représentative de

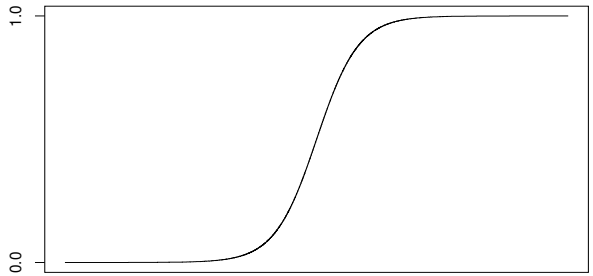
$$x \mapsto p_\beta(x) = \frac{\exp(x\beta)}{1+\exp(x\beta)}$$

pour $\beta = 0, 0.5, 2, 10$

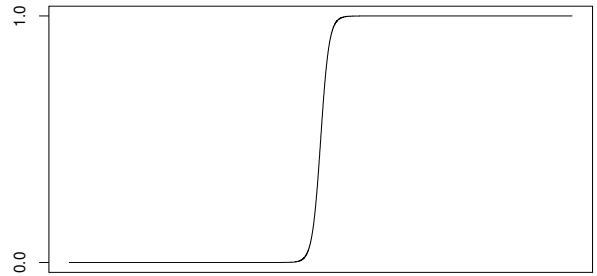
beta0



beta0.5



beta2



beta10

Lorsque β augmente, $p_\beta(x)$ est souvent proche de 0 ou 1.

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine*.
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{\frac{p_\beta(\tilde{x})}{1-p_\beta(\tilde{x})}}.$$

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine*.
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{\frac{p_\beta(\tilde{x})}{1-p_\beta(\tilde{x})}}.$$

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine*.
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{\frac{p_\beta(\tilde{x})}{1-p_\beta(\tilde{x})}}.$$

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine*.
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{\frac{p_\beta(\tilde{x})}{1-p_\beta(\tilde{x})}}.$$

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine*.
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{\frac{p_\beta(\tilde{x})}{1-p_\beta(\tilde{x})}}.$$

Interprétation des odds ratio

- Il faut être prudent avec l'interprétation des OR : ils sont très souvent utilisés mais pas toujours bien interprétés.

1. **Comparaison de probabilités de succès entre deux individus :**

$$\begin{array}{rcl} OR(x,\tilde{x}) > 1 & \Longleftrightarrow & p_\beta(x) > p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) = 1 & \Longleftrightarrow & p_\beta(x) = p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) < 1 & \Longleftrightarrow & p_\beta(x) < p_\beta(\tilde{x}) \end{array}$$

2. **Interprétation en termes de risque relatif :** dans le cas où $p(x)$ et $p(\tilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x,\tilde{x}) \approx p_\beta(x)/p_\beta(\tilde{x})$$

et interpréter "simplement".

Interprétation des odds ratio

- Il faut être prudent avec l'interprétation des OR : ils sont très souvent utilisés mais pas toujours bien interprétés.

1. **Comparaison de probabilités de succès entre deux individus** :

$$\begin{array}{lcl} OR(x,\tilde{x}) > 1 & \Longleftrightarrow & p_\beta(x) > p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) = 1 & \Longleftrightarrow & p_\beta(x) = p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) < 1 & \Longleftrightarrow & p_\beta(x) < p_\beta(\tilde{x}) \end{array}$$

2. **Interprétation en termes de risque relatif** : dans le cas où $p(x)$ et $p(\tilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x,\tilde{x}) \approx p_\beta(x)/p_\beta(\tilde{x})$$

et interpréter "simplement".

Interprétation des odds ratio

- Il faut être prudent avec l'interprétation des OR : ils sont très souvent utilisés mais pas toujours bien interprétés.

1. **Comparaison de probabilités de succès entre deux individus** :

$$\begin{array}{lcl} OR(x,\tilde{x}) > 1 & \Longleftrightarrow & p_\beta(x) > p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) = 1 & \Longleftrightarrow & p_\beta(x) = p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) < 1 & \Longleftrightarrow & p_\beta(x) < p_\beta(\tilde{x}) \end{array}$$

2. **Interprétation en termes de risque relatif** : dans le cas où $p(x)$ et $p(\tilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x,\tilde{x}) \approx p_\beta(x)/p_\beta(\tilde{x})$$

et interpréter "simplement".

3. **Mesure de l'impact d'une variable :** pour le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

il est facile de vérifier que

$$\text{OR}(x, \tilde{x}) = \exp(\beta_1(x_1 - \tilde{x}_1)) \ldots \exp(\beta_p(x_p - \tilde{x}_p)).$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et $\tilde{x}$ qui diffèrent uniquement par la $j$ème variable. On obtient alors

$$\text{OR}(x, \tilde{x}) = \exp(\beta_j(x_j - \tilde{x}_j)).$$

Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

3. **Mesure de l'impact d'une variable :** pour le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

il est facile de vérifier que

$$\text{OR}(x, \tilde{x}) = \exp(\beta_1(x_1 - \tilde{x}_1)) \ldots \exp(\beta_p(x_p - \tilde{x}_p)).$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et $\tilde{x}$ qui diffèrent uniquement par la $j$ème variable. On obtient alors

$$\text{OR}(x, \tilde{x}) = \exp(\beta_j(x_j - \tilde{x}_j)).$$

Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

1. Le modèle
   - Présentation
   - Identifiabilité et la matrice de design
   - Interprétation des coefficients

2. Estimation des paramètres
   - La vraisemblance
   - Existence et unicité de l'emv
   - L'algorithme IRLS
   - Comportement asymptotique de l'emv

3. Bibliographie

- On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables $X_1, \ldots, X_p$ défini par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

avec $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ qui contient les paramètres inconnus du modèle.

- Le caractère binaire de la variable à expliquer rend la méthode des moindres carrés impossible à mettre en oeuvre dans ce contexte.
- On rappelle que les estimateurs des moindres carrés du modèle linéaire gaussien coïncident avec les estimateurs du maximum de vraisemblance.
- C'est par cette approche que sont estimés les paramètres du modèle logistique à partir d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ (les variables aléatoires Y_i sont indépendantes).

- On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables $X_1, \ldots, X_p$ défini par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

avec $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ qui contient les paramètres inconnus du modèle.

- Le caractère binaire de la variable à expliquer rend la méthode des moindres carrés impossible à mettre en oeuvre dans ce contexte.

- On rappelle que les estimateurs des moindres carrés du modèle linéaire gaussien coïncident avec les estimateurs du maximum de vraisemblance.

- C'est par cette approche que sont estimés les paramètres du modèle logistique à partir d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ (les variables aléatoires Y_i sont indépendantes).

- On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables $X_1, \ldots, X_p$ défini par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

avec $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ qui contient les paramètres inconnus du modèle.

- Le caractère binaire de la variable à expliquer rend la méthode des moindres carrés impossible à mettre en oeuvre dans ce contexte.

- On rappelle que les estimateurs des moindres carrés du modèle linéaire gaussien coïncident avec les estimateurs du maximum de vraisemblance.

- C'est par cette approche que sont estimés les paramètres du modèle logistique à partir d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ (les variables aléatoires Y_i sont indépendantes).

- On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables $X_1, \ldots, X_p$ défini par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

avec $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ qui contient les paramètres inconnus du modèle.

- Le caractère binaire de la variable à expliquer rend la méthode des moindres carrés impossible à mettre en oeuvre dans ce contexte.

- On rappelle que les estimateurs des moindres carrés du modèle linéaire gaussien coïncident avec les estimateurs du maximum de vraisemblance.

- C'est par cette approche que sont estimés les paramètres du modèle logistique à partir d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ (les variables aléatoires Y_i sont indépendantes).

- Les variables aléatoires $Y_1, \ldots, Y_n$ étant discrètes et indépendantes, la vraisemblance du modèle logistique est définie par

$$L_n : \{0,1\}^n \times \mathbb{R}^p \to \mathbb{R}$$

$$(y_1, \ldots, y_n, \beta) \mapsto \prod_{i=1}^{n} \mathbf{P}_\beta(Y_i = y_i)$$

où $\mathbf{P}_\beta$ désigne la probabilité sous le modèle logistique de paramètre β.

- Pour simplifier, on notera $L_n(y_1, \ldots, y_n, \beta) = L_n(\beta)$ et $\mathcal{L}_n(\beta) = \log(L_n(\beta))$.

Propriété

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\}.$$

- Les variables aléatoires $Y_1, \ldots, Y_n$ étant discrètes et indépendantes, la vraisemblance du modèle logistique est définie par

$$L_n : \{0,1\}^n \times \mathbb{R}^p \to \mathbb{R}$$

$$(y_1, \ldots, y_n, \beta) \mapsto \prod_{i=1}^{n} \mathbf{P}_\beta(Y_i = y_i)$$

où $\mathbf{P}_\beta$ désigne la probabilité sous le modèle logistique de paramètre β.

- Pour simplifier, on notera $L_n(y_1, \ldots, y_n, \beta) = L_n(\beta)$ et $\mathcal{L}_n(\beta) = \log(L_n(\beta))$.

Propriété

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\}.$$

Calcul du gradient

- Un moyen naturel de maximiser la log-vraisemblance est d'annuler son gradient

$$\nabla \mathcal{L}_n(\beta) = \left(\frac{\partial \mathcal{L}_n}{\partial \beta_1}(\beta), \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}(\beta) \right).$$

- On montre que

$$\nabla \mathcal{L}_n(\beta) = \sum_{i=1}^{n} [x_i(y_i - p_\beta(x_i))] = \mathbb{X}'(\mathbb{Y} - \mathbb{P}_\beta)$$

avec

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad \mathbb{P}_\beta = \begin{pmatrix} p_\beta(x_1) \\ \vdots \\ p_\beta(x_n) \end{pmatrix}.$$

Calcul du gradient

- Un moyen naturel de maximiser la log-vraisemblance est d'annuler son gradient

$$\nabla \mathcal{L}_n(\beta) = \left(\frac{\partial \mathcal{L}_n}{\partial \beta_1}(\beta), \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}(\beta) \right).$$

- On montre que

$$\nabla \mathcal{L}_n(\beta) = \sum_{i=1}^{n} [x_i(y_i - p_\beta(x_i))] = \mathbb{X}'(\mathbb{Y} - \mathbb{P}_\beta)$$

avec

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad \mathbb{P}_\beta = \begin{pmatrix} p_\beta(x_1) \\ \vdots \\ p_\beta(x_n) \end{pmatrix}.$$

Equations du score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1, \ldots, p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Equations du score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1, \ldots, p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Equations du score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1, \ldots, p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Equations du score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1, \ldots, p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Equations du score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1, \ldots, p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Un premier résultat

Proposition

Soit $(x_i, y_i), i = 1, \ldots, n$ un nuage de points avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$. On suppose que la matrice de design $\mathbb{X}$ est de plein rang égal à p. Alors la log-vraisemblance

$$\mathbb{R}^p \to \mathbb{R} \tag{3}$$

$$\beta \mapsto \mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} \tag{4}$$

est strictement concave.

Conséquence importante

Un algorithme itératif convergera vers l'estimateur du maximum du vraisemblance lorsque celui-ci existe. Il n'y a pas de risque de tomber sur un maximum local.

Un premier résultat

Proposition

Soit $(x_i, y_i), i = 1, \ldots, n$ un nuage de points avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$. On suppose que la matrice de design $\mathbb{X}$ est de plein rang égal à p. Alors la log-vraisemblance

$$\mathbb{R}^p \to \mathbb{R} \tag{3}$$

$$\beta \mapsto \mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} \tag{4}$$

est strictement concave.

Conséquence importante

Un algorithme itératif convergera vers l'estimateur du maximum du vraisemblance lorsque celui-ci existe. Il n'y a pas de risque de tomber sur un maximum local.

Définition

On dit que l'estimateur du maximum de vraisemblance n'existe pas lorsque les équations de score n'admettent pas de solution finie.

- **Exemple :** On dispose d'un échantillon de taille $n = 200$:

	x_i	y_i
1	A	0
⋮	⋮	⋮
100	A	0
101	B	1
⋮	⋮	⋮
200	B	1

Table – Les 200 observations.

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x_i) = \beta_1 \mathbf{1}_{x_i=A} + \beta_2 \mathbf{1}_{x_i=B}$$

Définition

On dit que l'estimateur du maximum de vraisemblance n'existe pas lorsque les équations de score n'admettent pas de solution finie.

- **Exemple :** On dispose d'un échantillon de taille $n = 200$:

	x_i	y_i
1	A	0
⋮	⋮	⋮
100	A	0
101	B	1
⋮	⋮	⋮
200	B	1

Table – Les 200 observations.

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x_i) = \beta_1 \mathbf{1}_{x_i=A} + \beta_2 \mathbf{1}_{x_i=B}$$

- Il est facile de voir que lorsque $\beta_1 \to -\infty$ et $\beta_2 \to +\infty$, la vraisemblance $L_n(\beta)$ tend vers 1.
  - Les équations de score n'admettent pas de solution finie et l'emv n'existe pas.

```
> n <- 100
> X <- factor(c(rep("A",n),rep("B",n)))
> X <- c(rep(0,n),rep(1,n))
> Y <- factor(c(rep(0,n),rep(1,n)))
> model <- glm(Y~X,family=binomial)
Message d'avis :
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
```

  - Une alerte nous prévient que l'algorithme permettant d'estimer les paramètres n'a pas convergé. On peut vérifier cette convergence avec la commande :

```
> model$converged
[1] FALSE
```

- Il est facile de voir que lorsque $\beta_1 \to -\infty$ et $\beta_2 \to +\infty$, la vraisemblance $L_n(\beta)$ tend vers 1.
- Les équations de score n'admettent pas de solution finie et l'emv n'existe pas.

```
> n <- 100
> X <- factor(c(rep("A",n),rep("B",n)))
> X <- c(rep(0,n),rep(1,n))
> Y <- factor(c(rep(0,n),rep(1,n)))
> model <- glm(Y~X,family=binomial)
Message d'avis :
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
```

- Une alerte nous prévient que l'algorithme permettant d'estimer les paramètres n'a pas convergé. On peut vérifier cette convergence avec la commande :

```
> model$converged
[1] FALSE
```

- Il est facile de voir que lorsque $\beta_1 \to -\infty$ et $\beta_2 \to +\infty$, la vraisemblance $L_n(\beta)$ tend vers 1.
- Les équations de score n'admettent pas de solution finie et l'emv n'existe pas.

```
> n <- 100
> X <- factor(c(rep("A",n),rep("B",n)))
> X <- c(rep(0,n),rep(1,n))
> Y <- factor(c(rep(0,n),rep(1,n)))
> model <- glm(Y~X,family=binomial)
Message d'avis :
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
```

- Une alerte nous prévient que l'algorithme permettant d'estimer les paramètres n'a pas convergé. On peut vérifier cette convergence avec la commande :

```
> model$converged
[1] FALSE
```

- Il est facile de voir que lorsque $\beta_1 \to -\infty$ et $\beta_2 \to +\infty$, la vraisemblance $L_n(\beta)$ tend vers 1.
- Les équations de score n'admettent pas de solution finie et l'emv n'existe pas.

```
> n <- 100
> X <- factor(c(rep("A",n),rep("B",n)))
> X <- c(rep(0,n),rep(1,n))
> Y <- factor(c(rep(0,n),rep(1,n)))
> model <- glm(Y~X,family=binomial)
Message d'avis :
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
```

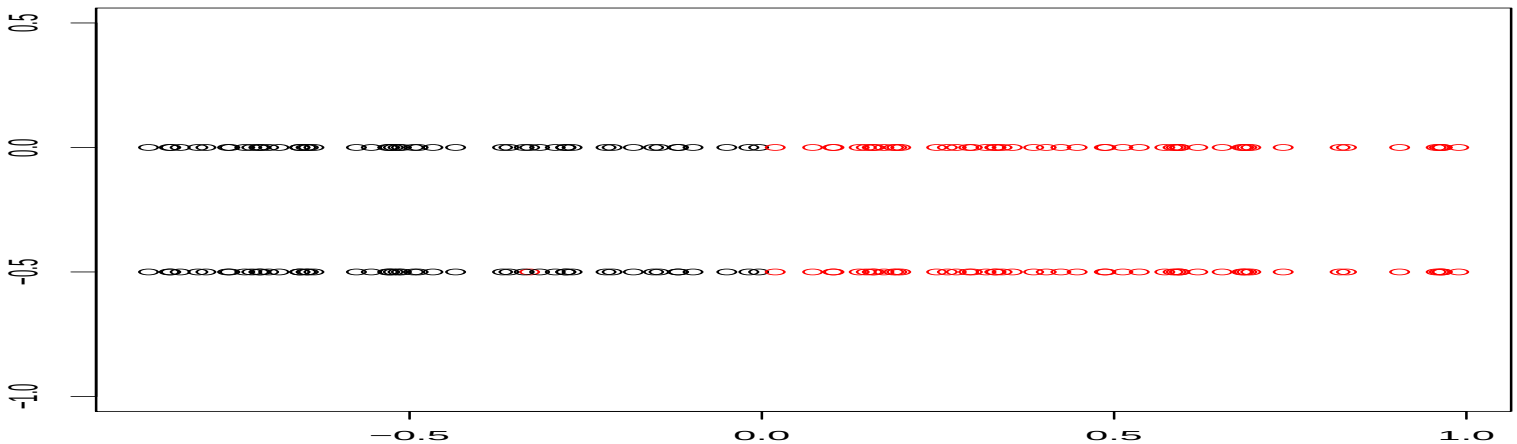
- Une alerte nous prévient que l'algorithme permettant d'estimer les paramètres n'a pas convergé. On peut vérifier cette convergence avec la commande :

```
> model$converged
[1] FALSE
```

Autre exemple

On considère 2 jeux de données générés selon le protocole suivant :

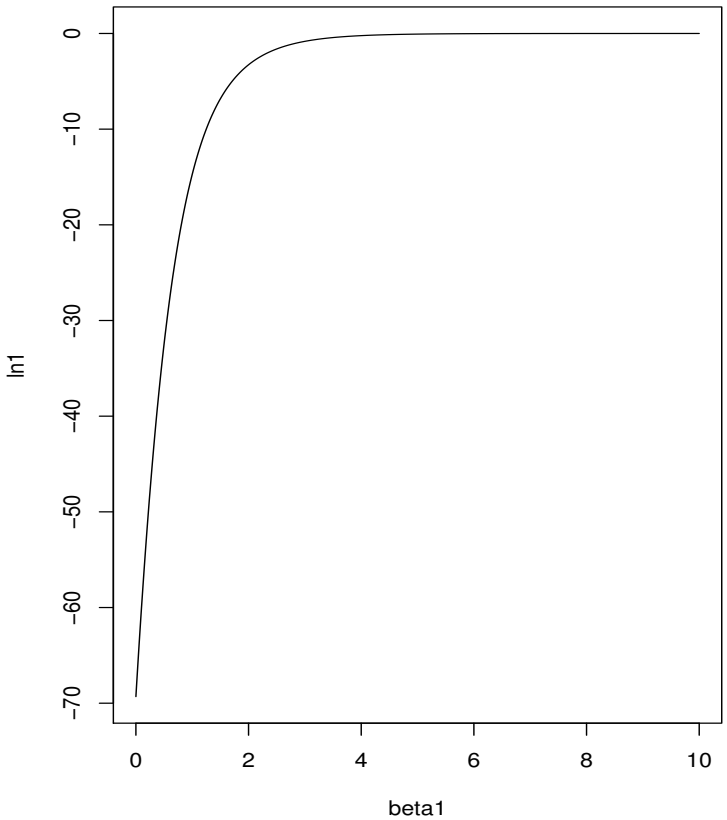
1. Le premier est tel que
   - pour $i = 1, \ldots, 50$, x_i est le réalisation d'une loi uniforme sur $[-1, 0]$ et $y_i = 0$;
   - pour $i = 51, \ldots, 100$, x_i est le réalisation d'une loi uniforme sur $[0, 1]$ et $y_i = 1$.
2. Le second nuage correspond au premier nuage dans lequel on a posé $y_1 = 1$.

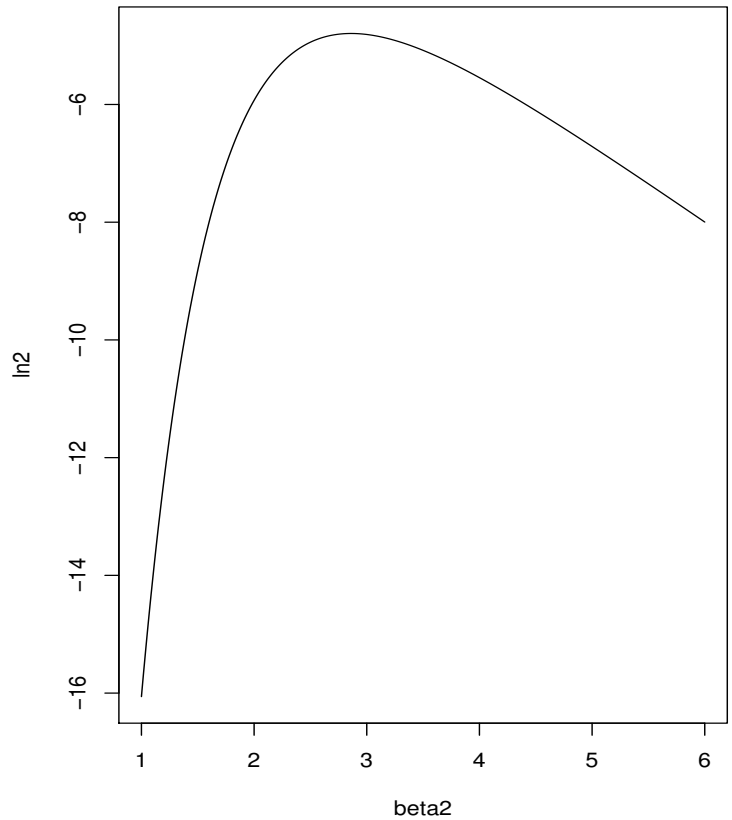


- On considère le modèle logistique

$$\text{logit}\, p_\beta(x_i) = \beta x_i.$$

- La figure suivante représente $\beta \mapsto \mathcal{L}_n(\beta)$ pour les deux jeux de données.





Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

- Les cas où l'estimation ne se passe pas bien ont une caractéristique commune : les modalités de Y sont parfaitement séparées selon les valeurs de X.
- Les problèmes d'estimation interviennent dans des situations similaires à celles-ci.
- Albert et Anderson (1984) ont précisé cette notion de séparabilité.

Définition

Un nuage de points $(x_1, y_1), \ldots, (x_n, y_n)$ avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ est dit :

- complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta > 0$ et $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta < 0$;
- quasi-complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta \geq 0$, $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta \leq 0$ et $\{i : x_i'\beta = 0\} \neq \emptyset$;
- en recouvrement s'il n'est ni complètement séparable ni quasi-complètement séparable.

- Les cas où l'estimation ne se passe pas bien ont une caractéristique commune : les modalités de Y sont parfaitement séparées selon les valeurs de X.
- Les problèmes d'estimation interviennent dans des situations similaires à celles-ci.
- Albert et Anderson (1984) ont précisé cette notion de séparabilité.

Définition

Un nuage de points $(x_1, y_1), \ldots, (x_n, y_n)$ avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ est dit :

- complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta > 0$ et $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta < 0$;
- quasi-complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta \geq 0$, $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta \leq 0$ et $\{i : x_i'\beta = 0\} \neq \emptyset$;
- en recouvrement s'il n'est ni complètement séparable ni quasi-complètement séparable.

- Les cas où l'estimation ne se passe pas bien ont une caractéristique commune : les modalités de Y sont parfaitement séparées selon les valeurs de X.
- Les problèmes d'estimation interviennent dans des situations similaires à celles-ci.
- Albert et Anderson (1984) ont précisé cette notion de séparabilité.

Définition

Un nuage de points $(x_1, y_1), \ldots, (x_n, y_n)$ avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ est dit :

- complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta > 0$ et $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta < 0$;
- quasi-complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta \geq 0$, $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta \leq 0$ et $\{i : x_i'\beta = 0\} \neq \emptyset$;
- en recouvrement s'il n'est ni complètement séparable ni quasi-complètement séparable.

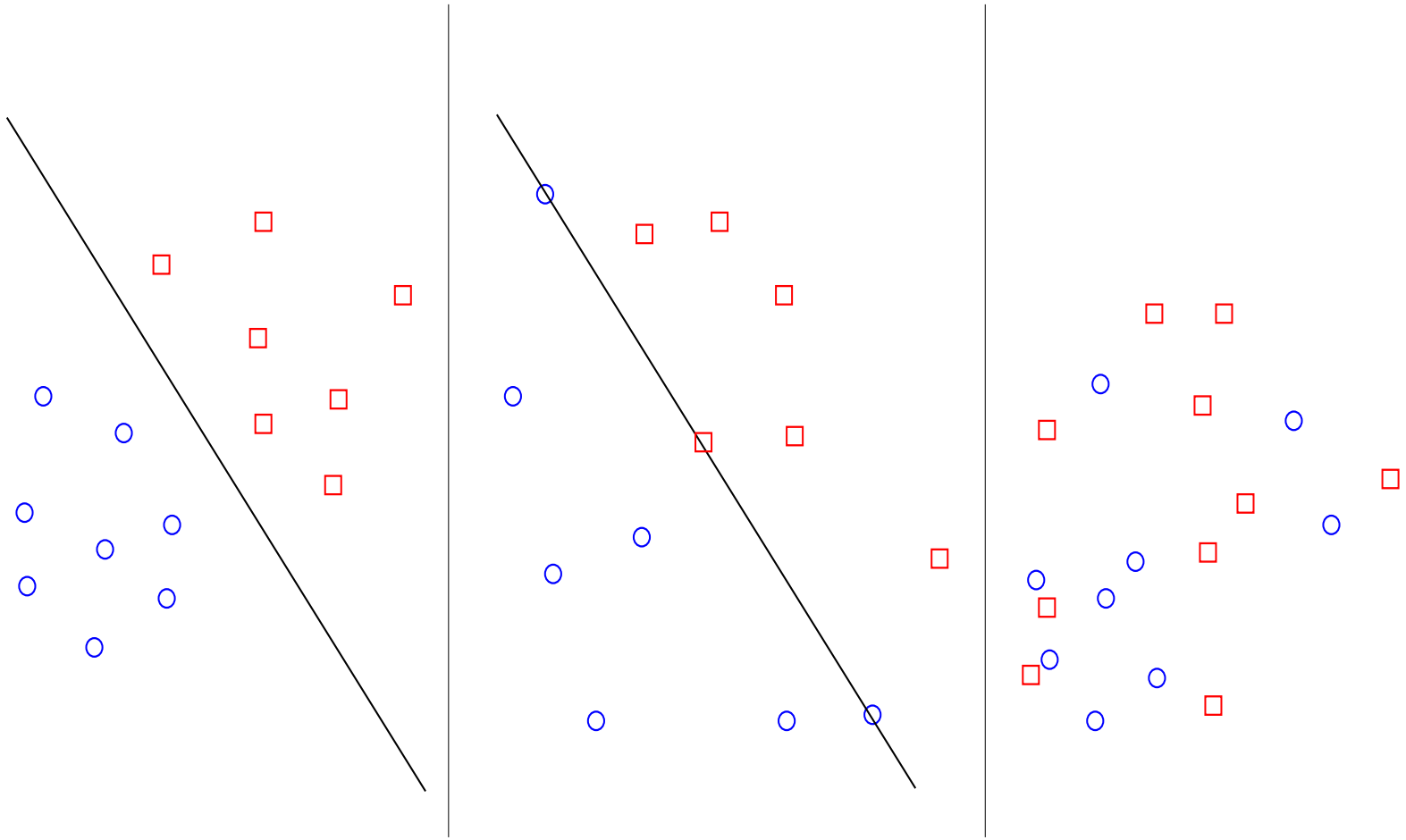


Figure – Exemple de séparabilité complète (gauche), quasi-complité (milieu) et de recouvrement (droite).

Existence de l'emv

Théorème [Albert and Anderson, 1984]

- Si le nuage de points est complètement séparable ou quasi-complètement séparable alors l'estimateur du maximum de vraisemblance n'existe pas.
- Si le nuage de points est en recouvrement alors l'estimateur du maximum de vraisemblance existe et est unique.

- Il est important de réaliser que dans la plupart des cas réels, les données ne sont pas séparées.
- Par conséquent, dans la plupart des cas réels, l'emv existe et est unique.
- Nécessité de trouver des algorithmes itératifs qui vont converger vers l'emv.

Existence de l'emv

Théorème [Albert and Anderson, 1984]

- Si le nuage de points est complètement séparable ou quasi-complètement séparable alors l'estimateur du maximum de vraisemblance n'existe pas.
- Si le nuage de points est en recouvrement alors l'estimateur du maximum de vraisemblance existe et est unique.

- Il est important de réaliser que dans la plupart des cas réels, les données ne sont pas séparées.
- Par conséquent, dans la plupart des cas réels, l'emv existe et est unique.
- Nécessité de trouver des algorithmes itératifs qui vont converger vers l'emv.

Existence de l'emv

Théorème [Albert and Anderson, 1984]

- Si le nuage de points est complètement séparable ou quasi-complètement séparable alors l'estimateur du maximum de vraisemblance n'existe pas.
- Si le nuage de points est en recouvrement alors l'estimateur du maximum de vraisemblance existe et est unique.

- Il est important de réaliser que dans la plupart des cas réels, les données ne sont pas séparées.
- Par conséquent, dans la plupart des cas réels, l'emv existe et est unique.
- Nécessité de trouver des algorithmes itératifs qui vont converger vers l'emv.

Existence de l'emv

Théorème [Albert and Anderson, 1984]

- Si le nuage de points est complètement séparable ou quasi-complètement séparable alors l'estimateur du maximum de vraisemblance n'existe pas.
- Si le nuage de points est en recouvrement alors l'estimateur du maximum de vraisemblance existe et est unique.

- Il est important de réaliser que dans la plupart des cas réels, les données ne sont pas séparées.
- Par conséquent, dans la plupart des cas réels, l'emv existe et est unique.
- Nécessité de trouver des algorithmes itératifs qui vont converger vers l'emv.

- L'approche consiste à trouver une suite $(\beta^{(k)})_{k\in\mathbb{N}}$ de vecteurs de $\mathbb{R}^p$ qui converge vers l'estimateur du maximum de vraisemblance $\hat{\beta}_n$.
- On rappelle que, si il existe, $\hat{\beta}_n$ est solution de l'équation de score et vérifie donc

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0. \tag{5}$$

- Soit $\beta^{(k)}$ un vecteur de $\mathbb{R}^p$. Un développement de Taylor à l'ordre 1 donne l'approximation

$$S(\hat{\beta}_n) \approx S(\beta^{(k)}) + A(\beta^{(k)})(\hat{\beta}_n - \beta^{(k)}) \tag{6}$$

où $A(\beta^{(k)})$ désigne la matrice héssienne de la log-vraisemblance au point $\beta^{(k)}$:

$$A(\beta^{(k)}) = \nabla^2 \mathcal{L}_n(\beta^{(k)}) = -\mathbb{X}' W_{\beta^{(k)}} \mathbb{X}$$

- Ici $W_{\beta^{(k)}}$ désigne la matrice diagonale $n \times n$ de terme général

$$p_{\beta^{(k)}}(x_i)(1 - p_{\beta^{(k)}}(x_i)), \quad i = 1, \ldots, n.$$

- L'approche consiste à trouver une suite $(\beta^{(k)})_{k\in\mathbb{N}}$ de vecteurs de $\mathbb{R}^p$ qui converge vers l'estimateur du maximum de vraisemblance $\hat{\beta}_n$.
- On rappelle que, si il existe, $\hat{\beta}_n$ est solution de l'équation de score et vérifie donc

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0. \tag{5}$$

- Soit $\beta^{(k)}$ un vecteur de $\mathbb{R}^p$. Un développement de Taylor à l'ordre 1 donne l'approximation

$$S(\hat{\beta}_n) \approx S(\beta^{(k)}) + A(\beta^{(k)})(\hat{\beta}_n - \beta^{(k)}) \tag{6}$$

où $A(\beta^{(k)})$ désigne la matrice héssienne de la log-vraisemblance au point $\beta^{(k)}$:

$$A(\beta^{(k)}) = \nabla^2 \mathcal{L}_n(\beta^{(k)}) = -\mathbb{X}' W_{\beta^{(k)}} \mathbb{X}$$

- Ici $W_{\beta^{(k)}}$ désigne la matrice diagonale $n \times n$ de terme général

$$p_{\beta^{(k)}}(x_i)(1 - p_{\beta^{(k)}}(x_i)), \quad i = 1, \ldots, n.$$

- Si $\mathbb{X}$ est de plein rang alors $A(\beta^{(k)})$ est inversible et on obtient en combinant (5) et (6)

$$\hat{\beta}_n \approx \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- Ce qui suggère d'utiliser la formule de récurrence

$$\beta^{(k+1)} = \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- D'où l'algorithme :

Algorithme IRLS

1. Initialisation : $\beta^{(0)}$, $k <- 1$
2. Répéter jusqu'à convergence ($\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$)
   1. $\beta^{(k+1)} <- \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)})$.
   2. $k <- k+1$

- Si $\mathbb{X}$ est de plein rang alors $A(\beta^{(k)})$ est inversible et on obtient en combinant (5) et (6)

$$\hat{\beta}_n \approx \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- Ce qui suggère d'utiliser la formule de récurrence

$$\beta^{(k+1)} = \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- D'où l'algorithme :

Algorithme IRLS

1. Initialisation : $\beta^{(0)}$, $k <- 1$
2. Répéter jusqu'à convergence ($\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$)
   1. $\beta^{(k+1)} <- \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)})$.
   2. $k <- k+1$

- Si $\mathbb{X}$ est de plein rang alors $A(\beta^{(k)})$ est inversible et on obtient en combinant (5) et (6)

$$\hat{\beta}_n \approx \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- Ce qui suggère d'utiliser la formule de récurrence

$$\beta^{(k+1)} = \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- D'où l'algorithme :

Algorithme IRLS

1. Initialisation : $\beta^{(0)}$, $k <- \quad 1$
2. Répéter jusqu'à convergence ($\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$)
   1. $\beta^{(k+1)} <- \quad \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)})$.
   2. $k <- \quad k+1$

Pourquoi IRLS ?

- La formule de récurrence de l'algorithme de maximisation peut se réécrire

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbf{X}' (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} (\mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} Z^{(k)},\end{aligned}$$

où $Z^{(k)} = \mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}})$.

- $\beta^{(k+1)}$ s'obtient en effectuant la régression pondérée du vecteur $Z^{(k)}$ par la matrice $\mathbb{X}$, d'où le nom de "Iterative Reweighted Least Square" (IRLS) pour cet algorithme.
- Les poids $W_{\beta^{(k)}}$ dépendent de $\mathbf{X}$ et $\beta^{(k)}$ et sont réévalués à chaque étape de l'algorithme.

Pourquoi IRLS ?

- La formule de récurrence de l'algorithme de maximisation peut se réécrire

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbf{X}' (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} (\mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} Z^{(k)},\end{aligned}$$

  où $Z^{(k)} = \mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}})$.

- $\beta^{(k+1)}$ s'obtient en effectuant la régression pondérée du vecteur $Z^{(k)}$ par la matrice $\mathbb{X}$, d'où le nom de "Iterative Reweighted Least Square" (IRLS) pour cet algorithme.
- Les poids $W_{\beta^{(k)}}$ dépendent de $\mathbf{X}$ et $\beta^{(k)}$ et sont réévalués à chaque étape de l'algorithme.

Pourquoi IRLS ?

- La formule de récurrence de l'algorithme de maximisation peut se réécrire

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbf{X}' (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} (\mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} Z^{(k)},\end{aligned}$$

  où $Z^{(k)} = \mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}})$.

- $\beta^{(k+1)}$ s'obtient en effectuant la régression pondérée du vecteur $Z^{(k)}$ par la matrice $\mathbb{X}$, d'où le nom de "Iterative Reweighted Least Square" (IRLS) pour cet algorithme.
- Les poids $W_{\beta^{(k)}}$ dépendent de $\mathbf{X}$ et $\beta^{(k)}$ et sont réévalués à chaque étape de l'algorithme.

Propriétés générales de l'emv

- N'ayant pas d'écriture explicite pour l'emv, il est "difficile" d'étudier les propriétés de l'emv pour le modèle logistique (contrairement au modèle linéaire gaussien).
- Néanmoins on sait que, sous certaines hypothèses de régularité, l'emv $\hat{\theta}_n$ d'un paramètre θ vérifie certaines propriétés asymptotiques :
  1. Consistance : $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta$ lorsque $n \to \infty$
  2. Normalité asymptotique :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}(\theta)^{-1})$$

où $\mathcal{I}(\theta)$ désigne la matrice d'information de Fisher du modèle au point θ.

Propriétés générales de l'emv

- N'ayant pas d'écriture explicite pour l'emv, il est "difficile" d'étudier les propriétés de l'emv pour le modèle logistique (contrairement au modèle linéaire gaussien).
- Néanmoins on sait que, sous certaines hypothèses de régularité, l'emv $\hat{\theta}_n$ d'un paramètre θ vérifie certaines propriétés asymptotiques :
  1. Consistance : $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta$ lorsque $n \to \infty$
  2. Normalité asymptotique :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}(\theta)^{-1})$$

où $\mathcal{I}(\theta)$ désigne la matrice d'information de Fisher du modèle au point θ.

Propriétés générales de l'emv

- N'ayant pas d'écriture explicite pour l'emv, il est "difficile" d'étudier les propriétés de l'emv pour le modèle logistique (contrairement au modèle linéaire gaussien).
- Néanmoins on sait que, sous certaines hypothèses de régularité, l'emv $\hat{\theta}_n$ d'un paramètre θ vérifie certaines propriétés asymptotiques :
  1. Consistance : $\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta$ lorsque $n \to \infty$
  2. Normalité asymptotique :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}(\theta)^{-1})$$

où $\mathcal{I}(\theta)$ désigne la matrice d'information de Fisher du modèle au point θ.

Comportement asymptotique de l'emv

Théorème [Fahrmeir and Kaufmann, 1985]

On suppose que :

- les $x_i, i = 1, \ldots, n$ prennent leurs valeurs dans une partie compacte de $\mathbb{R}^p$;
- la plus petite valeur propre $\lambda_{\min}(\mathbb{X}'\mathbb{X})$ tend vers $+\infty$ lorsque $n \to \infty$.

Alors

1. l'estimateur du maximum de vraisemblance existe asymptotiquement, c'est-à-dire qu'il existe une suite $\{\hat{\beta}_n\}_n$ de variables aléatoires

$$\lim_{n\to\infty} \mathbf{P}(S_n(\hat{\beta}_n) = 0) = 1.$$

2. la suite $\{\hat{\beta}_n\}_n$ est convergente : $\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta$.
3. $\{\hat{\beta}_n\}_n$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

Comportement asympotique de l'emv

Théorème [Fahrmeir and Kaufmann, 1985]

On suppose que :

- les $x_i, i = 1, \ldots, n$ prennent leurs valeurs dans une partie compacte de $\mathbb{R}^p$;
- la plus petite valeur propre $\lambda_{\min}(\mathbb{X}'\mathbb{X})$ tend vers $+\infty$ lorsque $n \to \infty$.

Alors

1. l'estimateur du maximum de vraisemblance existe asymptotiquement, c'est-à-dire qu'il existe une suite $\{\hat{\beta}_n\}_n$ de variables aléatoires

$$\lim_{n\to\infty} \mathbf{P}(S_n(\hat{\beta}_n) = 0) = 1.$$

2. la suite $\{\hat{\beta}_n\}_n$ est convergente : $\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta$.
3. $\{\hat{\beta}_n\}_n$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

Comportement asymptotique de l'emv

Théorème [Fahrmeir and Kaufmann, 1985]

On suppose que :

- les $x_i, i = 1, \ldots, n$ prennent leurs valeurs dans une partie compacte de $\mathbb{R}^p$;
- la plus petite valeur propre $\lambda_{\min}(\mathbb{X}'\mathbb{X})$ tend vers $+\infty$ lorsque $n \to \infty$.

Alors

1. l'estimateur du maximum de vraisemblance existe asymptotiquement, c'est-à-dire qu'il existe une suite $\{\hat{\beta}_n\}_n$ de variables aléatoires

$$\lim_{n\to\infty} \mathbf{P}(S_n(\hat{\beta}_n) = 0) = 1.$$

2. la suite $\{\hat{\beta}_n\}_n$ est convergente : $\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta$.
3. $\{\hat{\beta}_n\}_n$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

Comportement asymptotique de l'emv

Théorème [Fahrmeir and Kaufmann, 1985]

On suppose que :

- les $x_i, i = 1, \ldots, n$ prennent leurs valeurs dans une partie compacte de $\mathbb{R}^p$;
- la plus petite valeur propre $\lambda_{\min}(\mathbb{X}'\mathbb{X})$ tend vers $+\infty$ lorsque $n \to \infty$.

Alors

1. l'estimateur du maximum de vraisemblance existe asymptotiquement, c'est-à-dire qu'il existe une suite $\{\hat{\beta}_n\}_n$ de variables aléatoires

$$\lim_{n\to\infty} \mathbf{P}(S_n(\hat{\beta}_n) = 0) = 1.$$

2. la suite $\{\hat{\beta}_n\}_n$ est convergente : $\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta$.
3. $\{\hat{\beta}_n\}_n$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

Critique des hypothèses

Le théorème précédent repose sous deux hypothèses.

1. La compacité de l'espace des régresseurs n'est pas une hypothèse restrictive (En pratique, les valeurs des variables explicatives varient le plus souvent dans une partie compacte de $\mathbb{R}^p$).

2. La seconde hypothèse implique que l'information (au sens de Fisher) sur le paramètre β augmente lorsque le nombre d'observations tend vers $+\infty$. Elle est nécessaire pour augmenter la précision (en terme de diminution de la variance) de l'estimateur $\hat{\beta}_n$ lorsque le nombre d'observations augmente avec n.

Critique des hypothèses

Le théorème précédent repose sous deux hypothèses.

1. La compacité de l'espace des régresseurs n'est pas une hypothèse restrictive (En pratique, les valeurs des variables explicatives varient le plus souvent dans une partie compacte de $\mathbb{R}^p$).

2. La seconde hypothèse implique que l'information (au sens de Fisher) sur le paramètre β augmente lorsque le nombre d'observations tend vers $+\infty$. Elle est nécessaire pour augmenter la précision (en terme de diminution de la variance) de l'estimateur $\hat{\beta}_n$ lorsque le nombre d'observations augmente avec n.

Critique des hypothèses

Le théorème précédent repose sous deux hypothèses.

1. La compacité de l'espace des régresseurs n'est pas une hypothèse restrictive (En pratique, les valeurs des variables explicatives varient le plus souvent dans une partie compacte de $\mathbb{R}^p$).

2. La seconde hypothèse implique que l'information (au sens de Fisher) sur le paramètre β augmente lorsque le nombre d'observations tend vers $+\infty$. Elle est nécessaire pour augmenter la précision (en terme de diminution de la variance) de l'estimateur $\hat{\beta}_n$ lorsque le nombre d'observations augmente avec n.

Information de Fisher

- Le thérorème précédent n'est pas exploitable tel quel pour construire des intervalles de confiance ou des procédures de tests sur les paramètres du modèle (l'information de Fisher $\mathcal{I}(\beta)$ dépend du paramètre β qui est inconnu).
- On remarque que

$$(\hat{\beta}_n - \beta)' n\mathcal{I}(\beta)(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\mathcal{I}_n(\beta)(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

où $\mathcal{I}_n(\beta)$ désigne l'information de Fisher relative à l'ensemble des observations.

Propriété

$$\mathcal{I}_n(\beta) = -\mathbf{E}[\nabla^2 \mathcal{L}_n(\beta)] = \mathbb{X}' W_\beta \mathbb{X}.$$

Information de Fisher

- Le thérorème précédent n'est pas exploitable tel quel pour construire des intervalles de confiance ou des procédures de tests sur les paramètres du modèle (l'information de Fisher $\mathcal{I}(\beta)$ dépend du paramètre β qui est inconnu).
- On remarque que

$$(\hat{\beta}_n - \beta)' n\mathcal{I}(\beta)(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\mathcal{I}_n(\beta)(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

où $\mathcal{I}_n(\beta)$ désigne l'information de Fisher relative à l'ensemble des observations.

Propriété

$$\mathcal{I}_n(\beta) = -\mathbf{E}[\nabla^2 \mathcal{L}_n(\beta)] = \mathbb{X}' W_\beta \mathbb{X}.$$

Information de Fisher

- Le thérorème précédent n'est pas exploitable tel quel pour construire des intervalles de confiance ou des procédures de tests sur les paramètres du modèle (l'information de Fisher $\mathcal{I}(\beta)$ dépend du paramètre β qui est inconnu).
- On remarque que

$$(\hat{\beta}_n - \beta)' n\mathcal{I}(\beta)(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\mathcal{I}_n(\beta)(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

où $\mathcal{I}_n(\beta)$ désigne l'information de Fisher relative à l'ensemble des observations.

Propriété

$$\mathcal{I}_n(\beta) = -\mathbf{E}[\nabla^2 \mathcal{L}_n(\beta)] = \mathbb{X}' W_\beta \mathbb{X}.$$

Information de Fisher

- Le thérorème précédent n'est pas exploitable tel quel pour construire des intervalles de confiance ou des procédures de tests sur les paramètres du modèle (l'information de Fisher $\mathcal{I}(\beta)$ dépend du paramètre β qui est inconnu).
- On remarque que

$$(\hat{\beta}_n - \beta)' n\mathcal{I}(\beta)(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\mathcal{I}_n(\beta)(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

où $\mathcal{I}_n(\beta)$ désigne l'information de Fisher relative à l'ensemble des observations.

Propriété

$$\mathcal{I}_n(\beta) = -\mathbf{E}[\nabla^2 \mathcal{L}_n(\beta)] = \mathbb{X}' W_\beta \mathbb{X}.$$

Loi des estimateurs

- On estime $\mathcal{I}_n(\beta)$ par

$$\hat{\Sigma} = \mathcal{I}_n(\hat{\beta}_n) = \mathbb{X}' W_{\hat{\beta}_n} \mathbb{X}.$$

- On déduit de la convergence en probabilité de $\hat{\beta}_n$ vers β et des opérations classiques sur la convergence en loi (Slutsky) que

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

Propriété

On désigne par $\hat{\sigma}_j^2$ je $j$ème terme de la diagonale de $\hat{\Sigma}^{-1}$. On a alors

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi^2_1 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Loi des estimateurs

- On estime $\mathcal{I}_n(\beta)$ par

$$\hat{\Sigma} = \mathcal{I}_n(\hat{\beta}_n) = \mathbb{X}' W_{\hat{\beta}_n} \mathbb{X}.$$

- On déduit de la convergence en probabilité de $\hat{\beta}_n$ vers β et des opérations classiques sur la convergence en loi (Slutsky) que

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

Propriété

On désigne par $\hat{\sigma}_j^2$ je $j$ème terme de la diagonale de $\hat{\Sigma}^{-1}$. On a alors

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi^2_1 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Loi des estimateurs

- On estime $\mathcal{I}_n(\beta)$ par

$$\hat{\Sigma} = \mathcal{I}_n(\hat{\beta}_n) = \mathbb{X}' W_{\hat{\beta}_n} \mathbb{X}.$$

- On déduit de la convergence en probabilité de $\hat{\beta}_n$ vers β et des opérations classiques sur la convergence en loi (Slutsky) que

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

Propriété

On désigne par $\hat{\sigma}_j^2$ je $j$ème terme de la diagonale de $\hat{\Sigma}^{-1}$. On a alors

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi^2_1 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Intervalles de confiance et tests

1. Intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$IC_{1-\alpha}(\beta_j) = [\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j + u_{1-\alpha/2}\hat{\sigma}_j]$$

où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. Test (asymptotique) de nullité d'un paramètre au niveau α :
   - $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.
   - Sous H_0, $T = \hat{\beta}_j / \hat{\sigma}_j$ suit (approximativement) une loi $\mathcal{N}(0,1)$.
   - On rejette H_0 si $T_{obs} > u_{1-\alpha/2}$.

Intervalles de confiance et tests

1. Intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$IC_{1-\alpha}(\beta_j) = [\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j + u_{1-\alpha/2}\hat{\sigma}_j]$$

où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. Test (asymptotique) de nullité d'un paramètre au niveau α :
   - $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.
   - Sous H_0, $T = \hat{\beta}_j / \hat{\sigma}_j$ suit (approximativement) une loi $\mathcal{N}(0,1)$.
   - On rejette H_0 si $T_{obs} > u_{1-\alpha/2}$.

1. Intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$IC_{1-\alpha}(\beta_j) = [\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j + u_{1-\alpha/2}\hat{\sigma}_j]$$

où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. Test (asymptotique) de nullité d'un paramètre au niveau α :
   - $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.
   - Sous H_0, $T = \hat{\beta}_j/\hat{\sigma}_j$ suit (approximativement) une loi $\mathcal{N}(0,1)$.
   - On rejette H_0 si $T_{obs} > u_{1-\alpha/2}$.

Exemple

On reprend les données sur les pannes de machines.

```
> model <- glm(etat~.,data=panne,family=binomial)
```

- On obtient les tests de nullité des paramètres avec la fonction summary :

```
> summary(model)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.47808    0.83301   0.574    0.566
age          0.01388    0.09398   0.148    0.883
marqueB     -0.41941    0.81428  -0.515    0.607
marqueC     -1.45608    1.05358  -1.382    0.167
```

- Pour les intervalles de confiance, on utilise confint :

```
> confint(model)
                2.5 %    97.5 %
(Intercept) -1.1418097 2.2222689
age         -0.1721209 0.2086368
marqueB     -2.0793170 1.1657128
marqueC     -3.7421379 0.5176220
```

Exemple

On reprend les données sur les pannes de machines.

```
> model <- glm(etat~.,data=panne,family=binomial)
```

- On obtient les tests de nullité des paramètres avec la fonction summary :

```
> summary(model)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.47808    0.83301   0.574    0.566
age          0.01388    0.09398   0.148    0.883
marqueB     -0.41941    0.81428  -0.515    0.607
marqueC     -1.45608    1.05358  -1.382    0.167
```

- Pour les intervalles de confiance, on utilise confint :

```
> confint(model)
                 2.5 %    97.5 %
(Intercept) -1.1418097 2.2222689
age         -0.1721209 0.2086368
marqueB     -2.0793170 1.1657128
marqueC     -3.7421379 0.5176220
```

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Test de Wald

- Il est basé sur le résultat :

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

- On désigne par $\hat{\beta}_n^{(q)}$ les q premières composantes de $\hat{\beta}_n$ et $\hat{\Sigma}^{(q)}$ la matrice $q \times q$ comprenant les q premières lignes et colonnes de $\hat{\Sigma}$. On a alors :

$$(\hat{\beta}_n^{(q)} - \beta^{(q)})' \hat{\Sigma}^{(q)} (\hat{\beta}_n^{(q)} - \beta^{(q)}) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On déduit que sous H_0

$$\hat{\beta}^{(q)} \hat{\Sigma}^{(q)} \hat{\beta}^{(q)} \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Test de Wald

- Il est basé sur le résultat :

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

- On désigne par $\hat{\beta}_n^{(q)}$ les q premières composantes de $\hat{\beta}_n$ et $\hat{\Sigma}^{(q)}$ la matrice $q \times q$ comprenant les q premières lignes et colonnes de $\hat{\Sigma}$. On a alors :

$$(\hat{\beta}_n^{(q)} - \beta^{(q)})' \hat{\Sigma}^{(q)} (\hat{\beta}_n^{(q)} - \beta^{(q)}) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On déduit que sous H_0

$$\hat{\beta}^{(q)} \hat{\Sigma}^{(q)} \hat{\beta}^{(q)} \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Test de Wald

- Il est basé sur le résultat :

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

- On désigne par $\hat{\beta}_n^{(q)}$ les q premières composantes de $\hat{\beta}_n$ et $\hat{\Sigma}^{(q)}$ la matrice $q \times q$ comprenant les q premières lignes et colonnes de $\hat{\Sigma}$. On a alors :

$$(\hat{\beta}_n^{(q)} - \beta^{(q)})' \hat{\Sigma}^{(q)} (\hat{\beta}_n^{(q)} - \beta^{(q)}) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On déduit que sous H_0

$$\hat{\beta}^{(q)} \hat{\Sigma}^{(q)} \hat{\beta}^{(q)} \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Test de Wald

- Il est basé sur le résultat :

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

- On désigne par $\hat{\beta}_n^{(q)}$ les q premières composantes de $\hat{\beta}_n$ et $\hat{\Sigma}^{(q)}$ la matrice $q \times q$ comprenant les q premières lignes et colonnes de $\hat{\Sigma}$. On a alors :

$$(\hat{\beta}_n^{(q)} - \beta^{(q)})' \hat{\Sigma}^{(q)} (\hat{\beta}_n^{(q)} - \beta^{(q)}) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On déduit que sous H_0

$$\hat{\beta}^{(q)} \hat{\Sigma}^{(q)} \hat{\beta}^{(q)} \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Test de déviance ou du rapport de vraisemblance

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir

$$\hat{\beta}_{H_0} \approx \hat{\beta}_n \quad \text{et} \quad \mathcal{L}_n(\hat{\beta}_{H_0}) \approx \mathcal{L}_n(\hat{\beta}_n).$$

- Plus précisément, on montre que sous H_0,

$$2(\mathcal{L}_n(\hat{\beta}_n) - \mathcal{L}_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Test de déviance ou du rapport de vraisemblance

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir

$$\hat{\beta}_{H_0} \approx \hat{\beta}_n \quad \text{et} \quad \mathcal{L}_n(\hat{\beta}_{H_0}) \approx \mathcal{L}_n(\hat{\beta}_n).$$

- Plus précisément, on montre que sous H_0,

$$2(\mathcal{L}_n(\hat{\beta}_n) - \mathcal{L}_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Test de déviance ou du rapport de vraisemblance

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir

$$\hat{\beta}_{H_0} \approx \hat{\beta}_n \quad \text{et} \quad \mathcal{L}_n(\hat{\beta}_{H_0}) \approx \mathcal{L}_n(\hat{\beta}_n).$$

- Plus précisément, on montre que sous H_0,

$$2(\mathcal{L}_n(\hat{\beta}_n) - \mathcal{L}_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Test du score

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir $S(\hat{\beta}_{H_0}) = \nabla \mathcal{L}_n(\hat{\beta}_0) \approx 0$.

- Plus précisément, on montre que sous H_0,

$$S(\hat{\beta}_{H_0})' \hat{\Sigma}_{H_0}^{-1} S(\hat{\beta}_{H_0}) \xrightarrow{\mathcal{L}} \chi^2_q,$$

où $\hat{\Sigma}_{H_0} = \mathbb{X} W_{\hat{\beta}_{H_0}} \mathbb{X}$.

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Test du score

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir $S(\hat{\beta}_{H_0}) = \nabla \mathcal{L}_n(\hat{\beta}_0) \approx 0$.

- Plus précisément, on montre que sous H_0,

$$S(\hat{\beta}_{H_0})' \hat{\Sigma}_{H_0}^{-1} S(\hat{\beta}_{H_0}) \xrightarrow{\mathcal{L}} \chi^2_q,$$

où $\hat{\Sigma}_{H_0} = \mathbb{X} W_{\hat{\beta}_{H_0}} \mathbb{X}$.

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Test du score

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir $S(\hat{\beta}_{H_0}) = \nabla \mathcal{L}_n(\hat{\beta}_0) \approx 0$.

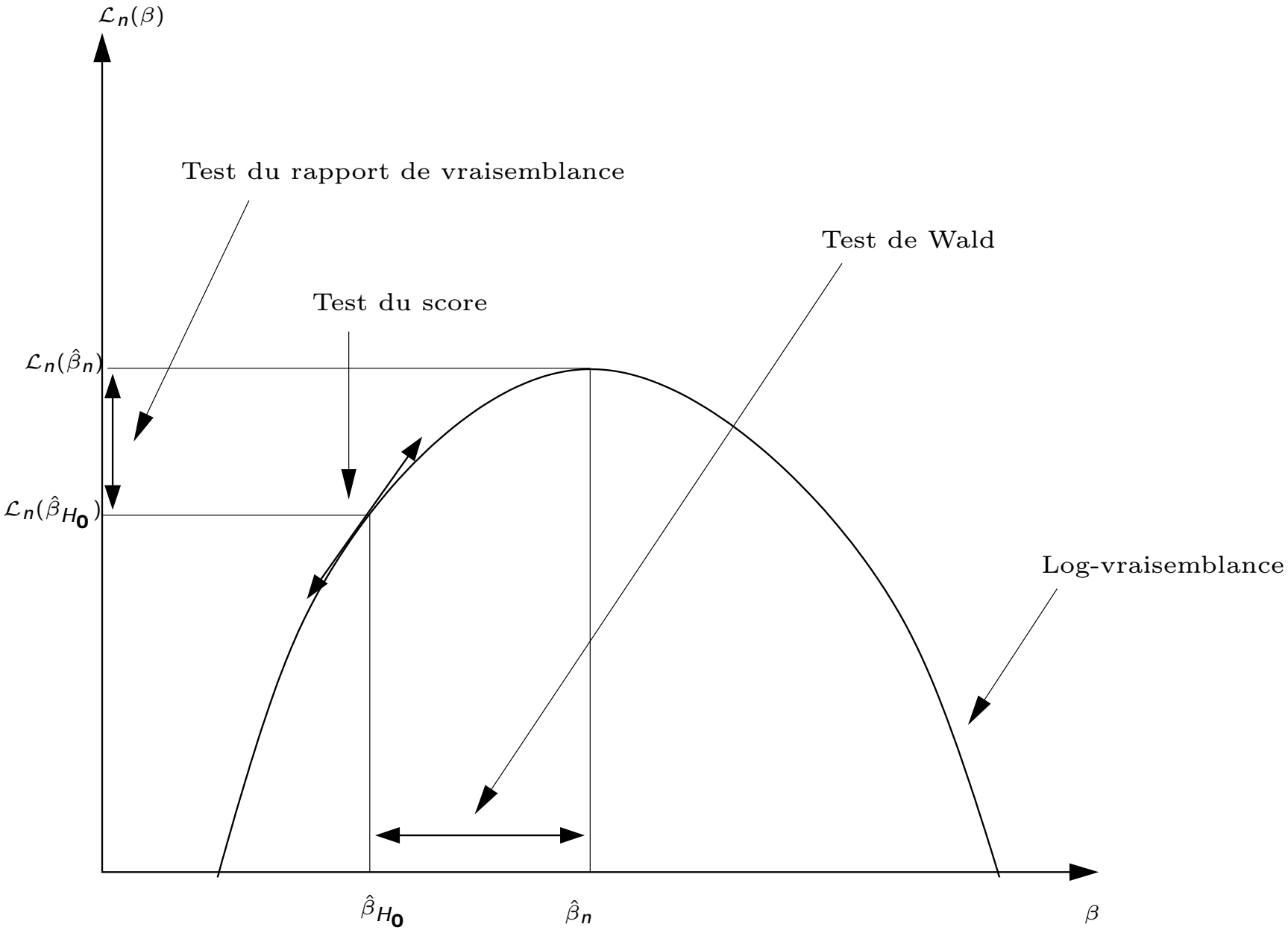
- Plus précisément, on montre que sous H_0,

$$S(\hat{\beta}_{H_0})' \hat{\Sigma}_{H_0}^{-1} S(\hat{\beta}_{H_0}) \xrightarrow{\mathcal{L}} \chi^2_q,$$

  où $\hat{\Sigma}_{H_0} = \mathbb{X} W_{\hat{\beta}_{H_0}} \mathbb{X}$.

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1-\alpha$ de la loi χ^2_q.

Récaptitulatif



Exemple sous R

- On peut tester l'effet des variables sour R avec la fonction Anova du package car :
  1. Pour le test de Wald :

```
> library(car)
> Anova(model,type=3,test.statistic="Wald")
Analysis of Deviance Table (Type III tests)

Response: etat
            Df  Chisq Pr(>Chisq)
(Intercept)  1 0.3294     0.5660
age          1 0.0218     0.8826
marque       2 1.9307     0.3809
Residuals   29
```

  2. Pour le test du rapport de vraisemblance :

```
> Anova(model,type=3,test.statistic="LR")
Analysis of Deviance Table (Type III tests)

Response: etat
       LR Chisq Df Pr(>Chisq)
age     0.02189  1     0.8824
marque  2.09562  2     0.3507
```

Exemple sous R

- On peut tester l'effet des variables sour R avec la fonction Anova du package car :

1. Pour le test de Wald :

```
> library(car)
> Anova(model,type=3,test.statistic="Wald")
Analysis of Deviance Table (Type III tests)

Response: etat
             Df  Chisq Pr(>Chisq)
(Intercept)   1 0.3294     0.5660
age           1 0.0218     0.8826
marque        2 1.9307     0.3809
Residuals    29
```

2. Pour le test du rapport de vraisemblance :

```
> Anova(model,type=3,test.statistic="LR")
Analysis of Deviance Table (Type III tests)

Response: etat
       LR Chisq Df Pr(>Chisq)
age     0.02189  1     0.8824
marque  2.09562  2     0.3507
```

Exemple sous R

- On peut tester l'effet des variables sour R avec la fonction Anova du package car :

1. Pour le test de Wald :

```
> library(car)
> Anova(model,type=3,test.statistic="Wald")
Analysis of Deviance Table (Type III tests)

Response: etat
             Df  Chisq Pr(>Chisq)
(Intercept)   1 0.3294     0.5660
age           1 0.0218     0.8826
marque        2 1.9307     0.3809
Residuals    29
```

2. Pour le test du rapport de vraisemblance :

```
> Anova(model,type=3,test.statistic="LR")
Analysis of Deviance Table (Type III tests)

Response: etat
       LR Chisq Df Pr(>Chisq)
age     0.02189  1     0.8824
marque  2.09562  2     0.3507
```

- Sous SAS, on utilise la `proc logistic`

```
proc logistic data=Tp1_panne descending;
class marque;
model panne= age marque;
run;
```

Le Système SAS

Procédure LOGISTIC

Statistiques d'ajustement du modèle		
Critère	**Constante uniquement**	**Constante et covariables**
AIC	47.717	51.502
SC	49.214	57.488
-2 Log	45.717	43.502

Test de l'hypothèse nulle globale : BETA=0			
Test	**Khi-2**	**DDL**	**Pr > Khi-2**
Rapport de vrais	2.2152	3	0.5290
Score	2.1630	3	0.5393
Wald	2.0333	3	0.5655

Analyse des effets Type 3			
Effet	**DDL**	**Khi-2 de Wald**	**Pr > Khi-2**
age	1	0.0218	0.8826
marque	2	1.9306	0.3809

Le Système SAS

Procédure LOGISTIC

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		**DDL**	**Valeur estimée**	**Erreur type**	**Khi-2 de Wald**	**Pr > Khi-2**
Intercept		1	-0.1471	0.6265	0.0551	0.8144
age		1	0.0139	0.0940	0.0218	0.8826
marque	**0**	1	0.6252	0.5344	1.3684	0.2421
marque	**1**	1	0.2058	0.4907	0.1758	0.6750

Estimations des rapports de cotes			
Effet	**Valeur estimée du point**	**Intervalle de confiance de Wald à 95 %**	
age	1.014	0.843	1.219
marque 0 vs 3	4.289	0.544	33.820
marque 1 vs 3	2.820	0.407	19.544

Références I


Albert, A. and Anderson, D. (1984).
On the existence of maximum likelihood estimates in logistic regression models.
Biometrika, 71 :1–10.


Fahrmeir, L. and Kaufmann, H. (1985).
Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models.
The Annals of Statistics, 13 :342–368.


Hosmer, D. and Lemeshow, S. (2000).
Applied Logistic Regression.
Wiley.

Troisième partie III

Sélection-"validation" de modèles

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

Pannes de machines

- Un chef d'entreprise souhaite vérifier la qualité d'un type de machines en fonction de l'âge et de la marque des moteurs. Il dispose
  1. d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
  2. d'une variable quantitative age repésentant l'âge du moteur ;
  3. d'une variable qualitative à 3 modalités `marque` représentant la marque du moteur,

- et de $n = 33$ observations :

```
> panne
   etat age marque
1     0   4      A
2     0   2      C
3     0   3      C
4     0   9      B
5     0   7      B
```

Pannes de machines

- Un chef d'entreprise souhaite vérifier la qualité d'un type de machines en fonction de l'âge et de la marque des moteurs. Il dispose
  1. d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
  2. d'une variable quantitative age repésentant l'âge du moteur ;
  3. d'une variable qualitative à 3 modalités `marque` représentant la marque du moteur,

- et de $n = 33$ observations :

```
> panne
   etat age marque
1     0   4      A
2     0   2      C
3     0   3      C
4     0   9      B
5     0   7      B
```

Role des femmes

- Il s'agit d'une étude effectuée en 1975 aux Etats-Unis. Il s'agit d'expliquer l'accord/désaccord d'individus avec la phrase

> Women should take care of running their homes and leave running the country up to men

par le sexe et le nombre d'années d'études des répondants.

```
> data("womensrole",package="HSAUR")
> womensrole <- womensrole[-24,]
> womensrole[1:5,]
  education  sex agree disagree
1         0 Male     4        2
2         1 Male     2        0
3         2 Male     4        0
4         3 Male     6        3
5         4 Male     5        5
```

Remarque

On est en présence de données répétées.

Role des femmes

- Il s'agit d'une étude effectuée en 1975 aux Etats-Unis. Il s'agit d'expliquer l'accord/désaccord d'individus avec la phrase

> Women should take care of running their homes and leave running the country up to men

par le sexe et le nombre d'années d'études des répondants.

```
> data("womensrole",package="HSAUR")
> womensrole <- womensrole[-24,]
> womensrole[1:5,]
  education  sex agree disagree
1         0 Male     4        2
2         1 Male     2        0
3         2 Male     4        0
4         3 Male     6        3
5         4 Male     5        5
```

Remarque

On est en présence de données répétées.

Role des femmes

- Il s'agit d'une étude effectuée en 1975 aux Etats-Unis. Il s'agit d'expliquer l'accord/désaccord d'individus avec la phrase

> Women should take care of running their homes and leave running the country up to men

par le sexe et le nombre d'années d'études des répondants.

```
> data("womensrole",package="HSAUR")
> womensrole <- womensrole[-24,]
> womensrole[1:5,]
  education  sex agree disagree
1         0 Male     4        2
2         1 Male     2        0
3         2 Male     4        0
4         3 Male     6        3
5         4 Male     5        5
```

Remarque

On est en présence de données répétées.

Maladie cardiovasculaire

- Il s'agit d'expliquer la présence/absence d'une maladie cardiovasculaire (chd) par 9 variables. On dispose de $n = 462$ individus.

```
> data(SAheart,package="bestglm")
> SAheart[1:5,]
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
```

Maladie cardiovasculaire

- Il s'agit d'expliquer la présence/absence d'une maladie cardiovasculaire (chd) par 9 variables. On dispose de $n = 462$ individus.

```
> data(SAheart,package="bestglm")
> SAheart[1:5,]
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
```

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

- Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
- On parlera toujours de meilleur modèle par rapport à un critère donné.
- On présentera deux types de critère :
  - ajustement du modèle (vraisemblances pénalisées)
  - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

   Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

Le problème de sélection/choix de modèles est ici abordé sous 2 angles (différents mais pas nécessairement indépendants) :

1. On est en présence de $\mathcal{M}_1, \ldots, \mathcal{M}_k$ modèles et on se pose le problème d'en choisir un.

   - Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle.
   - On parlera toujours de meilleur modèle par rapport à un critère donné.
   - On présentera deux types de critère :
     - ajustement du modèle (vraisemblances pénalisées)
     - capacité de prédiction du modèle

2. Etant données Y une variable à expliquer et $X_1, \ldots, X_p$ p variables explicatives, comment sélectionner automatiquement un modèle logistique ?

Il s'agit de trouver automatiquement un sous-groupe des variables explicatives permettant d'expliquer Y. On parle de sélection de variables.

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

Tests entre modèles emboités

- Afin de simplifier les notations, on supposera que l'on est en présence de deux modèles candidats $\mathcal{M}_1$ et $\mathcal{M}_2$.
- Nous nous plaçons dans le cas particulier où le modèle $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$ ($\mathcal{M}_1$ est un cas particulier de $\mathcal{M}_2$).

Exemple

$\mathcal{M}_1$ et $\mathcal{M}_2$ sont respectivement définis par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

et

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2.$$

Un moyen naturel de comparer $\mathcal{M}_1$ et $\mathcal{M}_2$ consiste à tester

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{contre} \quad H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0.$$

Tests entre modèles emboités

- Afin de simplifier les notations, on supposera que l'on est en présence de deux modèles candidats $\mathcal{M}_1$ et $\mathcal{M}_2$.
- Nous nous plaçons dans le cas particulier où le modèle $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$ ($\mathcal{M}_1$ est un cas particulier de $\mathcal{M}_2$).

Exemple

$\mathcal{M}_1$ et $\mathcal{M}_2$ sont respectivement définis par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

et

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2.$$

Un moyen naturel de comparer $\mathcal{M}_1$ et $\mathcal{M}_2$ consiste à tester

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{contre} \quad H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0.$$

Tests entre modèles emboités

- Afin de simplifier les notations, on supposera que l'on est en présence de deux modèles candidats $\mathcal{M}_1$ et $\mathcal{M}_2$.
- Nous nous plaçons dans le cas particulier où le modèle $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$ ($\mathcal{M}_1$ est un cas particulier de $\mathcal{M}_2$).

Exemple

$\mathcal{M}_1$ et $\mathcal{M}_2$ sont respectivement définis par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

et

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2.$$

Un moyen naturel de comparer $\mathcal{M}_1$ et $\mathcal{M}_2$ consiste à tester

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{contre} \quad H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0.$$

Tests entre modèles emboités

- Afin de simplifier les notations, on supposera que l'on est en présence de deux modèles candidats $\mathcal{M}_1$ et $\mathcal{M}_2$.
- Nous nous plaçons dans le cas particulier où le modèle $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$ ($\mathcal{M}_1$ est un cas particulier de $\mathcal{M}_2$).

Exemple

$\mathcal{M}_1$ et $\mathcal{M}_2$ sont respectivement définis par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

et

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \beta_2 x_2.$$

Un moyen naturel de comparer $\mathcal{M}_1$ et $\mathcal{M}_2$ consiste à tester

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{contre} \quad H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0.$$

- Plus généralement, considérons $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles logistiques à p_1 et p_2 paramètres tels que $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$.
- Tester $\mathcal{M}_1$ contre $\mathcal{M}_2$ revient à tester la nullité des coefficients de $\mathcal{M}_2$ que ne sont pas dans $\mathcal{M}_1$.
- On sait faire... On peut mettre en oeuvre un test de Wald, du rapport de vraisemblance ou du score.
- Sous H_0 ces 3 statistiques de test suivent une loi du $\chi^2_{p_2-p_1}$.

- Plus généralement, considérons $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles logistiques à p_1 et p_2 paramètres tels que $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$.
- Tester $\mathcal{M}_1$ contre $\mathcal{M}_2$ revient à tester la nullité des coefficients de $\mathcal{M}_2$ que ne sont pas dans $\mathcal{M}_1$.
- On sait faire... On peut mettre en oeuvre un test de Wald, du rapport de vraisemblance ou du score.
- Sous H_0 ces 3 statistiques de test suivent une loi du $\chi^2_{p_2-p_1}$.

- Plus généralement, considérons $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles logistiques à p_1 et p_2 paramètres tels que $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$.
- Tester $\mathcal{M}_1$ contre $\mathcal{M}_2$ revient à tester la nullité des coefficients de $\mathcal{M}_2$ que ne sont pas dans $\mathcal{M}_1$.
- On sait faire... On peut mettre en oeuvre un test de Wald, du rapport de vraisemblance ou du score.
- Sous H_0 ces 3 statistiques de test suivent une loi du $\chi^2_{p_2-p_1}$.

- Plus généralement, considérons $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles logistiques à p_1 et p_2 paramètres tels que $\mathcal{M}_1$ est emboité dans $\mathcal{M}_2$.
- Tester $\mathcal{M}_1$ contre $\mathcal{M}_2$ revient à tester la nullité des coefficients de $\mathcal{M}_2$ que ne sont pas dans $\mathcal{M}_1$.
- On sait faire... On peut mettre en oeuvre un test de Wald, du rapport de vraisemblance ou du score.
- Sous H_0 ces 3 statistiques de test suivent une loi du $\chi^2_{p_2-p_1}$.

Exemple

- Pour le problème sur la maladie cardiovasculaire, on souhaite comparer les modèles

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                    family=binomial)
```

  à l'aide d'un test de rapport de vraisemblance.

- On peut calculer la probabilité critique à la main

```
> stat <- 2*(logLik(model2)-logLik(model1))
> stat[1]
[1] 11.11016
> 1-pchisq(stat,df=length(model2$coef)-length(model1$coef))
[1] 0.003867751
```

- Ou directement avec la fonction anova

```
> anova(model1,model2,test="LRT")
Analysis of Deviance Table

Model 1: chd ~ tobacco + famhist
Model 2: chd ~ tobacco + famhist + adiposity + alcohol
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       459     524.58
2       457     513.47  2    11.11 0.003868 **
```

Exemple

- Pour le problème sur la maladie cardiovasculaire, on souhaite comparer les modèles

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                    family=binomial)
```

  à l'aide d'un test de rapport de vraisemblance.

- On peut calculer la probabilité critique à la main

```
> stat <- 2*(logLik(model2)-logLik(model1))
> stat[1]
[1] 11.11016
> 1-pchisq(stat,df=length(model2$coef)-length(model1$coef))
[1] 0.003867751
```

- Ou directement avec la fonction anova

```
> anova(model1,model2,test="LRT")
Analysis of Deviance Table

Model 1: chd ~ tobacco + famhist
Model 2: chd ~ tobacco + famhist + adiposity + alcohol
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       459     524.58
2       457     513.47  2    11.11 0.003868 **
```

Exemple

- Pour le problème sur la maladie cardiovasculaire, on souhaite comparer les modèles

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                    family=binomial)
```

  à l'aide d'un test de rapport de vraisemblance.

- On peut calculer la probabilité critique à la main

```
> stat <- 2*(logLik(model2)-logLik(model1))
> stat[1]
[1] 11.11016
> 1-pchisq(stat,df=length(model2$coef)-length(model1$coef))
[1] 0.003867751
```

- Ou directement avec la fonction anova

```
> anova(model1,model2,test="LRT")
Analysis of Deviance Table

Model 1: chd ~ tobacco + famhist
Model 2: chd ~ tobacco + famhist + adiposity + alcohol
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       459     524.58
2       457     513.47  2    11.11 0.003868 **
```

AIC-BIC

- Idée : utiliser la vraisemblance pour comparer $\mathcal{M}_1$ et $\mathcal{M}_2$.

Problème

Si $\mathcal{M}_1 \subset \mathcal{M}_2$ alors $\mathcal{L}_n(\hat{\beta}_1) \leq \mathcal{L}_n(\hat{\beta}_2)$ où $\hat{\beta}_j$ désigne l'emv du modèle $\mathcal{M}_j, j = 1, 2$.

- Conséquence : la vraisemblance sélectionnera toujours le modèle le plus complexe.

Solution

Pénaliser la vraisemblance par la complexité du modèle.

AIC-BIC

- Idée : utiliser la vraisemblance pour comparer $\mathcal{M}_1$ et $\mathcal{M}_2$.

Problème

Si $\mathcal{M}_1 \subset \mathcal{M}_2$ alors $\mathcal{L}_n(\hat{\beta}_1) \leq \mathcal{L}_n(\hat{\beta}_2)$ où $\hat{\beta}_j$ désigne l'emv du modèle $\mathcal{M}_j, j = 1, 2$.

- Conséquence : la vraisemblance sélectionnera toujours le modèle le plus complexe.

Solution

Pénaliser la vraisemblance par la complexité du modèle.

AIC-BIC

- Idée : utiliser la vraisemblance pour comparer $\mathcal{M}_1$ et $\mathcal{M}_2$.

Problème

Si $\mathcal{M}_1 \subset \mathcal{M}_2$ alors $\mathcal{L}_n(\hat{\beta}_1) \leq \mathcal{L}_n(\hat{\beta}_2)$ où $\hat{\beta}_j$ désigne l'emv du modèle $\mathcal{M}_j, j = 1, 2$.

- Conséquence : la vraisemblance sélectionnera toujours le modèle le plus complexe.

Solution

Pénaliser la vraisemblance par la complexité du modèle.

AIC-BIC

- Idée : utiliser la vraisemblance pour comparer $\mathcal{M}_1$ et $\mathcal{M}_2$.

Problème

Si $\mathcal{M}_1 \subset \mathcal{M}_2$ alors $\mathcal{L}_n(\hat{\beta}_1) \leq \mathcal{L}_n(\hat{\beta}_2)$ où $\hat{\beta}_j$ désigne l'emv du modèle $\mathcal{M}_j, j = 1, 2$.

- Conséquence : la vraisemblance sélectionnera toujours le modèle le plus complexe.

Solution

Pénaliser la vraisemblance par la complexité du modèle.

Définition

Soit $\mathcal{M}$ un modèle logistique à p paramètres. On note $\hat{\beta}_n$ l'emv des paramètres du modèle.

- L'AIC (Akaike Information Criterion) du modèle $\mathcal{M}$ est défini par

$$AIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + 2p.$$

- Le BIC (Bayesian Information Criterion) du modèle $\mathcal{M}$ est défini par

$$BIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + p \log n.$$

- Le modèle retenu sera celui qui minimise l'AIC ou le BIC.
- $\log n > 2$ (pour $n \geq 8$) BIC aura tendance a choisir des modèles plus parcimonieux que AIC.

Définition

Soit $\mathcal{M}$ un modèle logistique à p paramètres. On note $\hat{\beta}_n$ l'emv des paramètres du modèle.

- L'AIC (Akaike Information Criterion) du modèle $\mathcal{M}$ est défini par

$$AIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + 2p.$$

- Le BIC (Bayesian Information Criterion) du modèle $\mathcal{M}$ est défini par

$$BIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + p \log n.$$

- Le modèle retenu sera celui qui minimise l'AIC ou le BIC.
- $\log n > 2$ (pour $n \geq 8$) BIC aura tendance a choisir des modèles plus parcimonieux que AIC.

Définition

Soit $\mathcal{M}$ un modèle logistique à p paramètres. On note $\hat{\beta}_n$ l'emv des paramètres du modèle.

- L'AIC (Akaike Information Criterion) du modèle $\mathcal{M}$ est défini par

$$AIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + 2p.$$

- Le BIC (Bayesian Information Criterion) du modèle $\mathcal{M}$ est défini par

$$BIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\beta}_n) + p \log n.$$

- Le modèle retenu sera celui qui minimise l'AIC ou le BIC.
- $\log n > 2$ (pour $n \geq 8$) BIC aura tendance a choisir des modèles plus parcimonieux que AIC.

Exemple

- On considère les mêmes modèles que precédemment :

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                        family=binomial)
```

- On les compare en terme d'AIC et de BIC.

```
> c(AIC(model1),AIC(model2))
[1] 530.5759 523.4657
> c(BIC(model1),BIC(model2))
[1] 542.9826 544.1436
```

Conclusion

AIC sélectionne `model2` tandis que BIC sélectionne `model1`.

Exemple

- On considère les mêmes modèles que precédemment :

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                        family=binomial)
```

- On les compare en terme d'AIC et de BIC.

```
> c(AIC(model1),AIC(model2))
[1] 530.5759 523.4657
> c(BIC(model1),BIC(model2))
[1] 542.9826 544.1436
```

Conclusion

AIC sélectionne `model2` tandis que BIC sélectionne `model1`.

Exemple

- On considère les mêmes modèles que precédemment :

```
> model1 <- glm(chd~tobacco+famhist,data=SAheart,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=SAheart,
                                                        family=binomial)
```

- On les compare en terme d'AIC et de BIC.

```
> c(AIC(model1),AIC(model2))
[1] 530.5759 523.4657
> c(BIC(model1),BIC(model2))
[1] 542.9826 544.1436
```

Conclusion

AIC sélectionne `model2` tandis que BIC sélectionne `model1`.

Prévision

- L'idée est de chercher à comparer les pouvoirs de prédiction des modèles concurents et de choisir celui qui prédit le mieux.
- L'approche consiste à définir une règle de classification à partir d'un modèle logistique :

$$\hat{g} : \mathbb{R}^p \to \{0, 1\}$$

qui à une valeur observée des variables explicatives associe une valeur prédicte pour Y.

- Il existe plusieurs critères permettant de mesurer la performance d'une règle $\hat{g}$.
- Un des critère les plus classiques consiste à chercher à estimer la probabilité d'erreur

$$\mathbf{P}(\hat{g}(X) \neq Y).$$

Prévision

- L'idée est de chercher à comparer les pouvoirs de prédiction des modèles concurents et de choisir celui qui prédit le mieux.
- L'approche consiste à définir une règle de classification à partir d'un modèle logistique :

$$\hat{g} : \mathbb{R}^p \to \{0, 1\}$$

qui à une valeur observée des variables explicatives associe une valeur prédicte pour Y.

- Il existe plusieurs critères permettant de mesurer la performance d'une règle $\hat{g}$.
- Un des critère les plus classiques consiste à chercher à estimer la probabilité d'erreur

$$\mathbf{P}(\hat{g}(X) \neq Y).$$

Prévision

- L'idée est de chercher à comparer les pouvoirs de prédiction des modèles concurents et de choisir celui qui prédit le mieux.
- L'approche consiste à définir une règle de classification à partir d'un modèle logistique :

$$\hat{g} : \mathbb{R}^p \to \{0, 1\}$$

qui à une valeur observée des variables explicatives associe une valeur prédicte pour Y.

- Il existe plusieurs critères permettant de mesurer la performance d'une règle $\hat{g}$.
- Un des critère les plus classiques consiste à chercher à estimer la probabilité d'erreur

$$\mathbf{P}(\hat{g}(X) \neq Y).$$

Prévision avec un modèle logistique

- Modèle logistique permettant d'expliquer Y par $X_1, \ldots, X_p$:

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta.$$

- On peut estimer $p_\beta(x)$ par

$$p_{\hat{\beta}_n}(x) = \frac{\exp(x'\hat{\beta}_n)}{1 + \exp(x'\hat{\beta}_n)}.$$

- Un moyen naturel de prédire le label y_{n+1} d'un nouvel individu x_{n+1} est de poser

$$\hat{Y}_{n+1} = \begin{cases} 1 & \text{si } p_{\hat{\beta}_n}(x_{n+1}) \geq s \\ 0 & \text{sinon.} \end{cases}$$

Remarque

Le seuil s doit être choisi par l'utilisateur. Les logiciels prennent souvent par défaut $s = 0.5$.

Prévision avec un modèle logistique

- Modèle logistique permettant d'expliquer Y par $X_1, \ldots, X_p$:

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta.$$

- On peut estimer $p_\beta(x)$ par

$$p_{\hat{\beta}_n}(x) = \frac{\exp(x'\hat{\beta}_n)}{1 + \exp(x'\hat{\beta}_n)}.$$

- Un moyen naturel de prédire le label y_{n+1} d'un nouvel individu x_{n+1} est de poser

$$\hat{Y}_{n+1} = \begin{cases} 1 & \text{si } p_{\hat{\beta}_n}(x_{n+1}) \geq s \\ 0 & \text{sinon.} \end{cases}$$

Remarque

Le seuil s doit être choisi par l'utilisateur. Les logiciels prennent souvent par défaut $s = 0.5$.

Prévision avec un modèle logistique

- Modèle logistique permettant d'expliquer Y par $X_1, \ldots, X_p$:

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta.$$

- On peut estimer $p_\beta(x)$ par

$$p_{\hat{\beta}_n}(x) = \frac{\exp(x'\hat{\beta}_n)}{1 + \exp(x'\hat{\beta}_n)}.$$

- Un moyen naturel de prédire le label y_{n+1} d'un nouvel individu x_{n+1} est de poser

$$\hat{Y}_{n+1} = \begin{cases} 1 & \text{si } p_{\hat{\beta}_n}(x_{n+1}) \geq s \\ 0 & \textit{sinon.} \end{cases}$$

Remarque

Le seuil s doit être choisi par l'utilisateur. Les logiciels prennent souvent par défaut $s = 0.5$.

Prévision avec un modèle logistique

- Modèle logistique permettant d'expliquer Y par $X_1, \ldots, X_p$:

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta.$$

- On peut estimer $p_\beta(x)$ par

$$p_{\hat{\beta}_n}(x) = \frac{\exp(x'\hat{\beta}_n)}{1 + \exp(x'\hat{\beta}_n)}.$$

- Un moyen naturel de prédire le label y_{n+1} d'un nouvel individu x_{n+1} est de poser

$$\hat{Y}_{n+1} = \begin{cases} 1 & \text{si } p_{\hat{\beta}_n}(x_{n+1}) \geq s \\ 0 & \textit{sinon.} \end{cases}$$

Remarque

Le seuil s doit être choisi par l'utilisateur. Les logiciels prennent souvent par défaut $s = 0.5$.

Intervalles de confiance pour la probabilité estimée

- Etant donnée un nouvel individu x_{n+1}, il peut être intéressant de construire un intervalle de confiance pour la probabilité $p_\beta(x_{n+1})$.
- $\hat{\beta}_n$ est (pour n grand) un vecteur gaussien d'espérance β et de matrice de variance-covariance $\mathcal{I}_n(\beta)^{-1}$.
- Par conséquent, $x'_{n+1}\hat{\beta}_n$ est (pour n grand) une var de loi gaussienne

$$\mathcal{N}(x'_{n+1}\beta, x'_{n+1}\mathcal{I}_n(\beta)^{-1}x_{n+1}).$$

En posant $\hat{\sigma}^2 = x'_{n+1}(\mathbb{X}'W_{\hat{\beta}}\mathbb{X})^{-1}x_{n+1}$, on déduit

$$IC_{1-\alpha}(p_\beta(x_{n+1})) = \left[\frac{\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}; \frac{\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}\right]$$

- **Remarque** : il est possible de construire un IC centré en utilisant la delta-méthode.

Intervalles de confiance pour la probabilité estimée

- Etant donnée un nouvel individu x_{n+1}, il peut être intéressant de construire un intervalle de confiance pour la probabilité $p_\beta(x_{n+1})$.
- $\hat{\beta}_n$ est (pour n grand) un vecteur gaussien d'espérance β et de matrice de variance-covariance $\mathcal{I}_n(\beta)^{-1}$.
- Par conséquent, $x'_{n+1}\hat{\beta}_n$ est (pour n grand) une var de loi gaussienne

$$\mathcal{N}(x'_{n+1}\beta, x'_{n+1}\mathcal{I}_n(\beta)^{-1}x_{n+1}).$$

En posant $\hat{\sigma}^2 = x'_{n+1}(\mathbb{X}'W_{\hat{\beta}}\mathbb{X})^{-1}x_{n+1}$, on déduit

$$IC_{1-\alpha}(p_\beta(x_{n+1})) = \left[\frac{\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}; \frac{\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}\right]$$

- **Remarque** : il est possible de construire un IC centré en utilisant la delta-méthode.

Intervalles de confiance pour la probabilité estimée

- Etant donnée un nouvel individu x_{n+1}, il peut être intéressant de construire un intervalle de confiance pour la probabilité $p_\beta(x_{n+1})$.
- $\hat{\beta}_n$ est (pour n grand) un vecteur gaussien d'espérance β et de matrice de variance-covariance $\mathcal{I}_n(\beta)^{-1}$.
- Par conséquent, $x'_{n+1}\hat{\beta}_n$ est (pour n grand) une var de loi gaussienne

$$\mathcal{N}(x'_{n+1}\beta, x'_{n+1}\mathcal{I}_n(\beta)^{-1}x_{n+1}).$$

En posant $\hat{\sigma}^2 = x'_{n+1}(\mathbb{X}'W_{\hat{\beta}}\mathbb{X})^{-1}x_{n+1}$, on déduit

$$IC_{1-\alpha}(p_\beta(x_{n+1})) = \left[\frac{\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})} ; \frac{\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})} \right]$$

- **Remarque** : il est possible de construire un IC centré en utilisant la delta-méthode.

Intervalles de confiance pour la probabilité estimée

- Etant donnée un nouvel individu x_{n+1}, il peut être intéressant de construire un intervalle de confiance pour la probabilité $p_\beta(x_{n+1})$.
- $\hat{\beta}_n$ est (pour n grand) un vecteur gaussien d'espérance β et de matrice de variance-covariance $\mathcal{I}_n(\beta)^{-1}$.
- Par conséquent, $x'_{n+1}\hat{\beta}_n$ est (pour n grand) une var de loi gaussienne

$$\mathcal{N}(x'_{n+1}\beta, x'_{n+1}\mathcal{I}_n(\beta)^{-1}x_{n+1}).$$

En posant $\hat{\sigma}^2 = x'_{n+1}(\mathbb{X}'W_{\hat{\beta}}\mathbb{X})^{-1}x_{n+1}$, on déduit

$$IC_{1-\alpha}(p_\beta(x_{n+1})) = \left[\frac{\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n - u_{1-\alpha_2}\hat{\sigma})}; \frac{\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}{1+\exp(x'_{n+1}\hat{\beta}_n + u_{1-\alpha_2}\hat{\sigma})}\right]$$

- **Remarque** : il est possible de construire un IC centré en utilisant la delta-méthode.

Un critère de prévision : la probabilité d'erreur

- On suppose dans cette section que les variables explicatives sont aléatoires et on note $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ un n-échantillon i.i.d de même loi que (X, Y).
- L'approche consiste à comparer deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ en comparant les probabilités d'erreur

$$L(\hat{g}) = \mathbf{P}(\hat{g}(X) \neq Y)$$

des règles de classification issues de ces deux modèles.

La probabilité $L(\hat{g})$ est inconnue et doit être estimée.

Un critère de prévision : la probabilité d'erreur

- On suppose dans cette section que les variables explicatives sont aléatoires et on note $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ un n-échantillon i.i.d de même loi que (X, Y).
- L'approche consiste à comparer deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ en comparant les probabilités d'erreur

$$L(\hat{g}) = \mathbf{P}(\hat{g}(X) \neq Y)$$

des règles de classification issues de ces deux modèles.

La probabilité $L(\hat{g})$ est inconnue et doit être estimée.

Un critère de prévision : la probabilité d'erreur

- On suppose dans cette section que les variables explicatives sont aléatoires et on note $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ un n-échantillon i.i.d de même loi que (X, Y).
- L'approche consiste à comparer deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ en comparant les probabilités d'erreur

$$L(\hat{g}) = \mathbf{P}(\hat{g}(X) \neq Y)$$

des règles de classification issues de ces deux modèles.

La probabilité $L(\hat{g})$ est inconnue et doit être estimée.

- Première idée : estimer $L(\hat{g})$ en
  1. appliquant la règle $\hat{g}$ sur les variables X_i pour en déduire une prévision $\hat{Y}_i = \hat{g}(X_i)$ de la variable Y pour chaque individu.
  2. comparant la prévision $\hat{Y}_i$ avec la valeur observée Y_i

$$L_n(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{g}(X_i)\neq Y_i}.$$

Table de confusion

- On dresse généralement la table de confusion

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	OK	E_1
$Y=1$	E_2	OK

- La probabilité d'erreur est alors estimée par

$$L_n(\hat{g}) = \frac{E_1 + E_2}{n}.$$

- Première idée : estimer $L(\hat{g})$ en
  1. appliquant la règle $\hat{g}$ sur les variables X_i pour en déduire une prévision $\hat{Y}_i = \hat{g}(X_i)$ de la variable Y pour chaque individu.
  2. comparant la prévision $\hat{Y}_i$ avec la valeur observée Y_i

$$L_n(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{g}(X_i)\neq Y_i}.$$

Table de confusion

- On dresse généralement la table de confusion

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	OK	E_1
$Y=1$	E_2	OK

- La probabilité d'erreur est alors estimée par

$$L_n(\hat{g}) = \frac{E_1 + E_2}{n}.$$

- Première idée : estimer $L(\hat{g})$ en
  1. appliquant la règle $\hat{g}$ sur les variables X_i pour en déduire une prévision $\hat{Y}_i = \hat{g}(X_i)$ de la variable Y pour chaque individu.
  2. comparant la prévision $\hat{Y}_i$ avec la valeur observée Y_i

$$L_n(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{g}(X_i)\neq Y_i}.$$

Table de confusion

- On dresse généralement la table de confusion

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	OK	E_1
$Y=1$	E_2	OK

- La probabilité d'erreur est alors estimée par

$$L_n(\hat{g}) = \frac{E_1 + E_2}{n}.$$

- Première idée : estimer $L(\hat{g})$ en
  1. appliquant la règle $\hat{g}$ sur les variables X_i pour en déduire une prévision $\hat{Y}_i = \hat{g}(X_i)$ de la variable Y pour chaque individu.
  2. comparant la prévision $\hat{Y}_i$ avec la valeur observée Y_i

$$L_n(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{g}(X_i)\neq Y_i}.$$

Table de confusion

- On dresse généralement la table de confusion

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	OK	E_1
$Y=1$	E_2	OK

- La probabilité d'erreur est alors estimée par

$$L_n(\hat{g}) = \frac{E_1 + E_2}{n}.$$

Problème

- $L_n(\hat{g})$ n'est généralement pas un bon estimateur de $L(\hat{g})$ (sous-estimation).
- La loi des grands nombres ne peux s'appliquer car les variables

$$\mathbf{1}_{\hat{g}(X_i)\neq Y_i}$$

ne sont pas indépendantes.

- Le problème vient du fait que l'échantillon $\mathcal{D}_n$ est utilisé deux fois (pour calculer $\hat{g}$ puis pour estimer $L(\hat{g})$).

Solution

Découper l'échantillon en deux :

- un échantillon d'apprentissage utilisé pour calculer la règle $\hat{g}$ (estimer les paramètres du modèle logistique).
- un échantillon test ou de validation utilisé pour estimer la probabilité d'erreur $L(\hat{g})$.

Problème

- $L_n(\hat{g})$ n'est généralement pas un bon estimateur de $L(\hat{g})$ (sous-estimation).
- La loi des grands nombres ne peux s'appliquer car les variables

$$\mathbf{1}_{\hat{g}(X_i)\neq Y_i}$$

ne sont pas indépendantes.

- Le problème vient du fait que l'échantillon $\mathcal{D}_n$ est utilisé deux fois (pour calculer $\hat{g}$ puis pour estimer $L(\hat{g})$).

Solution

Découper l'échantillon en deux :

- un échantillon d'apprentissage utilisé pour calculer la règle $\hat{g}$ (estimer les paramètres du modèle logistique).
- un échantillon test ou de validation utilisé pour estimer la probabilité d'erreur $L(\hat{g})$.

La procédure A/V

Estimateur de $L(\hat{g})$ par A/V

L'échantillon $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ est séparé aléatoirement en deux sous échantillons :

1. un échantillon d'apprentissage $\mathcal{D}_\ell = \{(X_i, Y_i), i \in \mathcal{J}_\ell\}$ de taille ℓ utilisé pour estimer les paramètres du modèle et en déduire la règle $\hat{g}$.
2. un échantillon test ou de validation $\mathcal{D}_m = \{(X_i, Y_i), i \in \mathcal{J}_m\}$ de taille m utilisé pour estimer $L(\hat{g})$ par

$$L_n(\hat{g}) = \frac{1}{m} \sum_{i \in \mathcal{J}_m} \mathbf{1}_{\hat{g}(X_i) \neq Y_i},$$

avec $\mathcal{J}_\ell \cup \mathcal{J}_m = \{1, \ldots n\}$ et $\mathcal{J}_\ell \cap \mathcal{J}_m = \emptyset$.

Propriété

L'estimateur $L_n(\hat{g})$ est un estimateur sans biais de $L(\hat{g})$.

La procédure A/V

Estimateur de $L(\hat{g})$ par A/V

L'échantillon $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ est séparé aléatoirement en deux sous échantillons :

1. un échantillon d'apprentissage $\mathcal{D}_\ell = \{(X_i, Y_i), i \in \mathcal{J}_\ell\}$ de taille ℓ utilisé pour estimer les paramètres du modèle et en déduire la règle $\hat{g}$.
2. un échantillon test ou de validation $\mathcal{D}_m = \{(X_i, Y_i), i \in \mathcal{J}_m\}$ de taille m utilisé pour estimer $L(\hat{g})$ par

$$L_n(\hat{g}) = \frac{1}{m} \sum_{i \in \mathcal{J}_m} \mathbf{1}_{\hat{g}(X_i) \neq Y_i},$$

avec $\mathcal{J}_\ell \cup \mathcal{J}_m = \{1, \ldots n\}$ et $\mathcal{J}_\ell \cap \mathcal{J}_m = \emptyset$.

Propriété

L'estimateur $L_n(\hat{g})$ est un estimateur sans biais de $L(\hat{g})$.

Exemple

On estime la probabilité d'erreur pour deux modèles logistiques concurrents sur les données concernant la maladie cardiovasculaire.

- Construction des échantillons d'apprentissage et test

```
n <- nrow(SAheart)
l <- 250 #taille de l'ech d'apprentissage
set.seed(1234)
perm <- sample(n)
dapp <- SAheart[perm[1:l],]
dtest <- SAheart[-perm[1:l],]
```

- Ajustement des modèles sur l'échantillon d'apprentissage.

```
> model1 <- glm(chd~tobacco+famhist,data=dapp,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=dapp,family=binomia
```

- Estimation de la probabilité d'erreur sur l'échantillon test.

```
> prev1 <- round(predict(model1,newdata=dtest,type="response"))
> prev2 <- round(predict(model2,newdata=dtest,type="response"))
>
> mean(prev1!=dtest$chd)
[1] 0.3113208
> mean(prev2!=dtest$chd)
[1] 0.2877358
```

Exemple

On estime la probabilité d'erreur pour deux modèles logistiques concurrents sur les données concernant la maladie cardiovasculaire.

- Construction des échantillons d'apprentissage et test

```
n <- nrow(SAheart)
l <- 250 #taille de l'ech d'apprentissage
set.seed(1234)
perm <- sample(n)
dapp <- SAheart[perm[1:l],]
dtest <- SAheart[-perm[1:l],]
```

- Ajustement des modèles sur l'échantillon d'apprentissage.

```
> model1 <- glm(chd~tobacco+famhist,data=dapp,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=dapp,family=binomia
```

- Estimation de la probabilité d'erreur sur l'échantillon test.

```
> prev1 <- round(predict(model1,newdata=dtest,type="response"))
> prev2 <- round(predict(model2,newdata=dtest,type="response"))
>
> mean(prev1!=dtest$chd)
[1] 0.3113208
> mean(prev2!=dtest$chd)
[1] 0.2877358
```

Exemple

On estime la probabilité d'erreur pour deux modèles logistiques concurrents sur les données concernant la maladie cardiovasculaire.

- Construction des échantillons d'apprentissage et test

```
n <- nrow(SAheart)
l <- 250 #taille de l'ech d'apprentissage
set.seed(1234)
perm <- sample(n)
dapp <- SAheart[perm[1:l],]
dtest <- SAheart[-perm[1:l],]
```

- Ajustement des modèles sur l'échantillon d'apprentissage.

```
> model1 <- glm(chd~tobacco+famhist,data=dapp,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=dapp,family=binomia
```

- Estimation de la probabilité d'erreur sur l'échantillon test.

```
> prev1 <- round(predict(model1,newdata=dtest,type="response"))
> prev2 <- round(predict(model2,newdata=dtest,type="response"))
>
> mean(prev1!=dtest$chd)
[1] 0.3113208
> mean(prev2!=dtest$chd)
[1] 0.2877358
```

Exemple

On estime la probabilité d'erreur pour deux modèles logistiques concurrents sur les données concernant la maladie cardiovasculaire.

- Construction des échantillons d'apprentissage et test

```
n <- nrow(SAheart)
l <- 250 #taille de l'ech d'apprentissage
set.seed(1234)
perm <- sample(n)
dapp <- SAheart[perm[1:l],]
dtest <- SAheart[-perm[1:l],]
```

- Ajustement des modèles sur l'échantillon d'apprentissage.

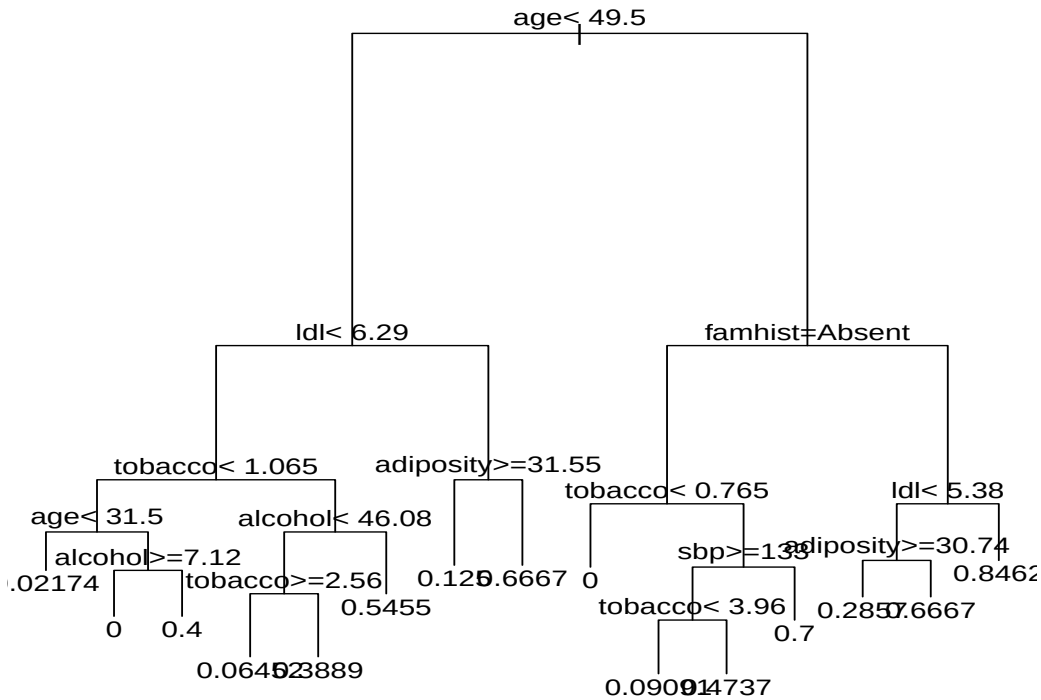
```
> model1 <- glm(chd~tobacco+famhist,data=dapp,family=binomial)
> model2 <- glm(chd~tobacco+famhist+adiposity+alcohol,data=dapp,family=binomia
```

- Estimation de la probabilité d'erreur sur l'échantillon test.

```
> prev1 <- round(predict(model1,newdata=dtest,type="response"))
> prev2 <- round(predict(model2,newdata=dtest,type="response"))
>
> mean(prev1!=dtest$chd)
[1] 0.3113208
> mean(prev2!=dtest$chd)
[1] 0.2877358
```

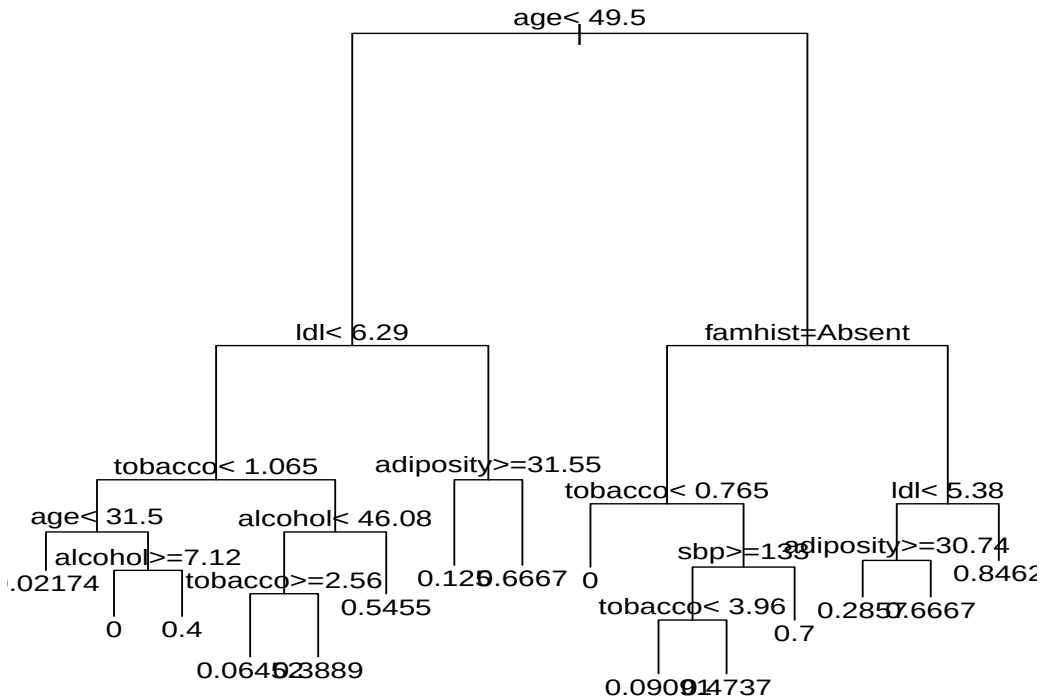
- Un des avantages de la probabilité d'erreur est qu'elle permet de comparer différents modèles issus de différentes méthodes.
- Construisons par exemple un arbre de classification.

```
> arbre <- rpart(chd~.,data=dapp)
> plot(arbre)
> text(arbre,pretty=0)
```



- Un des avantages de la probabilité d'erreur est qu'elle permet de comparer différents modèles issus de différentes méthodes.
- Construisons par exemple un arbre de classification.

```
> arbre <- rpart(chd~.,data=dapp)
> plot(arbre)
> text(arbre,pretty=0)
```



- Et estimons sa probabilité d'erreur sur l'échantillon test

```
> prev3 <- predict(arbre,newdata=dtest,type="class")
> mean(prev3!=dtest$chd)
[1] 0.3490566
```

Probabilités d'erreur estimées

modèle	logit1	logit2	arbre
erreur estimée	0.31	0.29	0.35

- Pour ce critère, on privilégiera le second modèle logistique.

- Et estimons sa probabilité d'erreur sur l'échantillon test

```
> prev3 <- predict(arbre,newdata=dtest,type="class")
> mean(prev3!=dtest$chd)
[1] 0.3490566
```

Probabilités d'erreur estimées

modèle	logit1	logit2	arbre
erreur estimée	0.31	0.29	0.35

- Pour ce critère, on privilégiera le second modèle logistique.

- Et estimons sa probabilité d'erreur sur l'échantillon test

```
> prev3 <- predict(arbre,newdata=dtest,type="class")
> mean(prev3!=dtest$chd)
[1] 0.3490566
```

Probabilités d'erreur estimées

modèle	logit1	logit2	arbre
erreur estimée	0.31	0.29	0.35

- Pour ce critère, on privilégiera le second modèle logistique.

Motivations

- Dans la partie précédente, on a présenté des outils permettant de comparer des modèles construits.
- On se place dans un cadre différent : étant donné p variables explicatives $X_1, \ldots, X_p$, on cherche une procédure automatique permettant de trouver le "meilleur" sous-groupe de variables à mettre dans le modèle logistique.

Pourquoi ?

(Au moins) 2 raisons peuvent motiver cette démarche :

1. **Descriptif** : identifier les variables qui permettent d'expliquer la cible.
2. **Statistique** : la variance des estimateurs augmente avec le nombre de paramètres du modèle. Diminuer le nombre de variables permettra d'avoir des estimateurs plus précis.

Motivations

- Dans la partie précédente, on a présenté des outils permettant de comparer des modèles construits.
- On se place dans un cadre différent : étant donné p variables explicatives $X_1, \ldots, X_p$, on cherche une procédure automatique permettant de trouver le "meilleur" sous-groupe de variables à mettre dans le modèle logistique.

Pourquoi ?

(Au moins) 2 raisons peuvent motiver cette démarche :

1. **Descriptif** : identifier les variables qui permettent d'expliquer la cible.
2. **Statistique** : la variance des estimateurs augmente avec le nombre de paramètres du modèle. Diminuer le nombre de variables permettra d'avoir des estimateurs plus précis.

Motivations

- Dans la partie précédente, on a présenté des outils permettant de comparer des modèles construits.
- On se place dans un cadre différent : étant donné p variables explicatives $X_1, \ldots, X_p$, on cherche une procédure automatique permettant de trouver le "meilleur" sous-groupe de variables à mettre dans le modèle logistique.

Pourquoi ?

(Au moins) 2 raisons peuvent motiver cette démarche :

1. **Descriptif** : identifier les variables qui permettent d'expliquer la cible.
2. **Statistique** : la variance des estimateurs augmente avec le nombre de paramètres du modèle. Diminuer le nombre de variables permettra d'avoir des estimateurs plus précis.

Motivations

- Dans la partie précédente, on a présenté des outils permettant de comparer des modèles construits.
- On se place dans un cadre différent : étant donné p variables explicatives $X_1, \ldots, X_p$, on cherche une procédure automatique permettant de trouver le "meilleur" sous-groupe de variables à mettre dans le modèle logistique.

Pourquoi ?

(Au moins) 2 raisons peuvent motiver cette démarche :

1. **Descriptif** : identifier les variables qui permettent d'expliquer la cible.
2. **Statistique** : la variance des estimateurs augmente avec le nombre de paramètres du modèle. Diminuer le nombre de variables permettra d'avoir des estimateurs plus précis.

Recherche exhaustive

- Une approche naturelle est de construire tous les modèles logistiques (2^p) et de retenir celui qui optimise un critère donné (AIC-BIC...).
- Les package leaps permet de faire cela pour la régression linéaire.
- Pour le modèle logistique, on peut utiliser le package bestglm.

```
> library(bestglm)
> model4 <- bestglm(dapp,family=binomial,IC="BIC")
Morgan-Tatar search since family is non-gaussian.
> model4$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
   (Intercept)            ldl  famhistPresent            age
      -4.29645        0.18650        0.82172        0.05088

Degrees of Freedom: 249 Total (i.e. Null);  246 Residual
Null Deviance:     319.2
Residual Deviance: 267.5  AIC: 275.5
```

Recherche exhaustive

- Une approche naturelle est de construire tous les modèles logistiques (2^p) et de retenir celui qui optimise un critère donné (AIC-BIC...).
- Les package leaps permet de faire cela pour la régression linéaire.
- Pour le modèle logistique, on peut utiliser le package bestglm.

```
> library(bestglm)
> model4 <- bestglm(dapp,family=binomial,IC="BIC")
Morgan-Tatar search since family is non-gaussian.
> model4$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
   (Intercept)            ldl  famhistPresent            age
      -4.29645        0.18650        0.82172        0.05088

Degrees of Freedom: 249 Total (i.e. Null);  246 Residual
Null Deviance:     319.2
Residual Deviance: 267.5  AIC: 275.5
```

Recherche exhaustive

- Une approche naturelle est de construire tous les modèles logistiques (2^p) et de retenir celui qui optimise un critère donné (AIC-BIC...).
- Les package leaps permet de faire cela pour la régression linéaire.
- Pour le modèle logistique, on peut utiliser le package bestglm.

```
> library(bestglm)
> model4 <- bestglm(dapp,family=binomial,IC="BIC")
Morgan-Tatar search since family is non-gaussian.
> model4$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
   (Intercept)            ldl  famhistPresent            age
      -4.29645        0.18650        0.82172        0.05088

Degrees of Freedom: 249 Total (i.e. Null);  246 Residual
Null Deviance:     319.2
Residual Deviance: 267.5  AIC: 275.5
```

Recherche exhaustive

- Une approche naturelle est de construire tous les modèles logistiques (2^p) et de retenir celui qui optimise un critère donné (AIC-BIC...).
- Les package leaps permet de faire cela pour la régression linéaire.
- Pour le modèle logistique, on peut utiliser le package bestglm.

```
> library(bestglm)
> model4 <- bestglm(dapp,family=binomial,IC="BIC")
Morgan-Tatar search since family is non-gaussian.
> model4$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
   (Intercept)             ldl  famhistPresent             age
      -4.29645         0.18650         0.82172         0.05088

Degrees of Freedom: 249 Total (i.e. Null);  246 Residual
Null Deviance:     319.2
Residual Deviance: 267.5  AIC: 275.5
```

- On peut également visualiser les variables retenues dans les meilleurs modèles pour le critère donné

```
> model4$BestModels
    sbp tobacco   ldl adiposity famhist typea obesity alcohol  age Criterion
1 FALSE   FALSE  TRUE      FALSE    TRUE FALSE   FALSE   FALSE TRUE  284.0427
2 FALSE   FALSE  TRUE      FALSE   FALSE FALSE   FALSE   FALSE TRUE  286.0520
3 FALSE    TRUE  TRUE      FALSE    TRUE FALSE   FALSE   FALSE TRUE  286.7856
4 FALSE   FALSE FALSE      FALSE    TRUE FALSE   FALSE   FALSE TRUE  287.3270
5 FALSE   FALSE  TRUE      FALSE    TRUE  TRUE   FALSE   FALSE TRUE  287.9329
```

Lorsque le nombre de variables p est trop grand, balayer tous les modèles peut se révéler très couteux en tant de calcul. On a alors recours à des méthodes pas à pas.

Méthode pas à pas

L'approche consiste à :

- construire un modèle initial
- Ajouter (forward) ou supprimer (backward) la variable qui optimise un critère donné (BIC ou AIC) par exemple.
- Répéter le processus jusqu'à un critère d'arrêt.

Méthode pas à pas

L'approche consiste à :

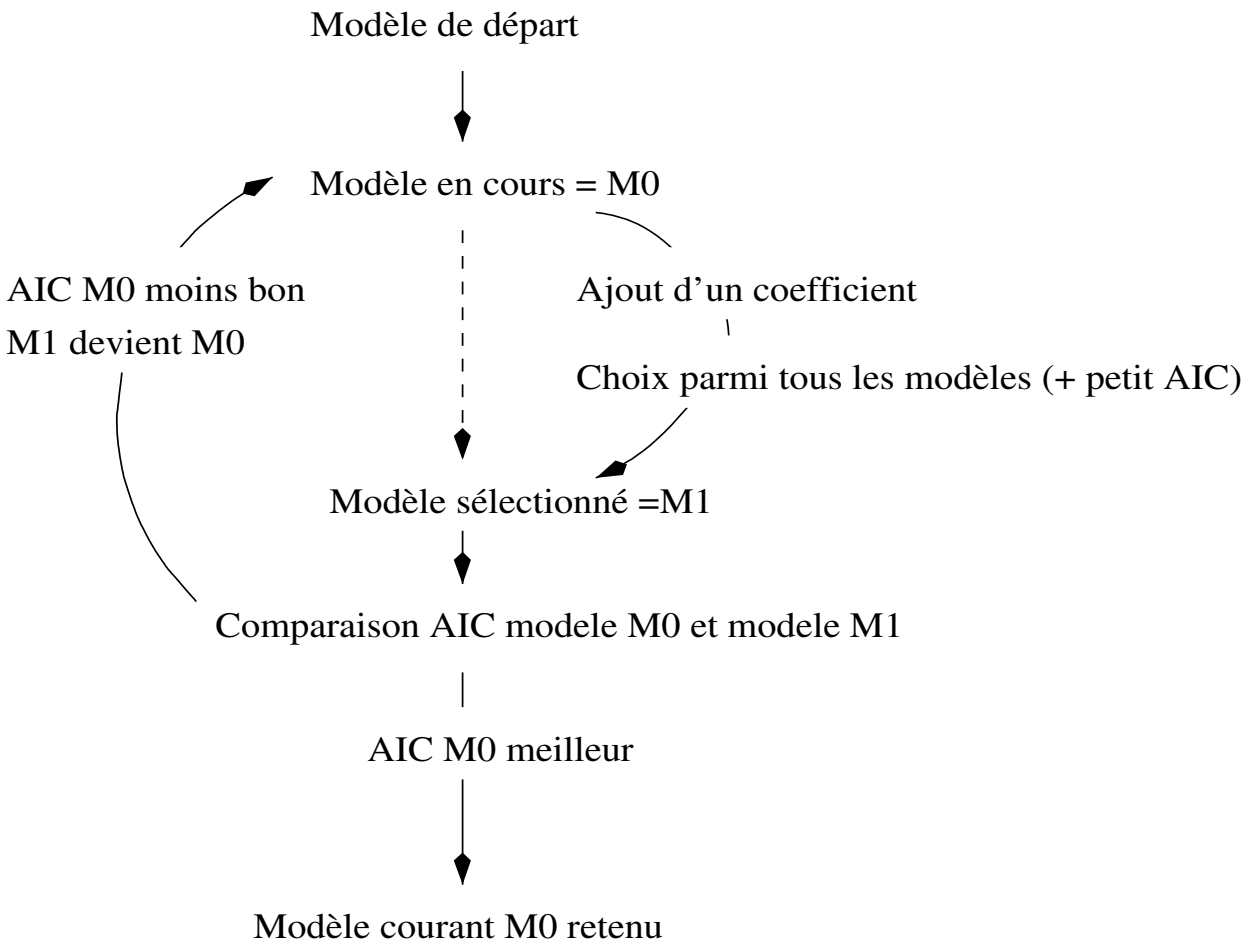
- construire un modèle initial
- Ajouter (forward) ou supprimer (backward) la variable qui optimise un critère donné (BIC ou AIC) par exemple.
- Répéter le processus jusqu'à un critère d'arrêt.

Méthode pas à pas

L'approche consiste à :

- construire un modèle initial
- Ajouter (forward) ou supprimer (backward) la variable qui optimise un critère donné (BIC ou AIC) par exemple.
- Répéter le processus jusqu'à un critère d'arrêt.

Technique ascendante utilisant l'AIC



Exemple sur R

- La fonction step permet de sélectionner des variables à l'aide de méthodes pas à pas.

```
> model_complet <- glm(chd~.,data=dapp,family=binomial)
> model_step <- step(model_complet,direction="backward",k=log(nrow(dapp)))
> model_step

Call:  glm(formula = chd ~ ldl + famhist + age, family = binomial,
                                                    data = dapp)

Coefficients:
   (Intercept)             ldl  famhistPresent             age
      -4.29645         0.18650         0.82172         0.05088
```

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

- Poser un modèle revient à faire une hypothèse : la loi de la variable d'intérêt appartient à une famille de loi donnée.
- Pour le modèle logistique cette hypothèse est que la loi des Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

- Les résultats présentés précédemment sont vrais uniquement sous cette hypothèse. Il faut par conséquent la vérifier.

Les techniques permettant (dans une certaine mesure) de vérifier cette hypothèse sont similaires à celles du modèle de régression linéaire (tests d'adéquation, étude des résidus...).

- Poser un modèle revient à faire une hypothèse : la loi de la variable d'intérêt appartient à une famille de loi donnée.
- Pour le modèle logistique cette hypothèse est que la loi des Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

- Les résultats présentés précédemment sont vrais uniquement sous cette hypothèse. Il faut par conséquent la vérifier.

Les techniques permettant (dans une certaine mesure) de vérifier cette hypothèse sont similaires à celles du modèle de régression linéaire (tests d'adéquation, étude des résidus...).

- Poser un modèle revient à faire une hypothèse : la loi de la variable d'intérêt appartient à une famille de loi donnée.
- Pour le modèle logistique cette hypothèse est que la loi des Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

- Les résultats présentés précédemment sont vrais uniquement sous cette hypothèse. Il faut par conséquent la vérifier.

Les techniques permettant (dans une certaine mesure) de vérifier cette hypothèse sont similaires à celles du modèle de régression linéaire (tests d'adéquation, étude des résidus...).

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

La déviance

- Idée : se baser sur la vraisemblance. En effet, plus la vraisamblance est proche de 1, plus le modèle est "proche" des données.
- La valeur d'une vraisemblance est difficile à interpréter (elle dépend notamment du nombre de données).
- La déviance permet de comparer la vraisemblance du modèle à celle d'un modèle parfait en terme d'adequation aux données : le modèle saturé.

La déviance

- Idée : se baser sur la vraisemblance. En effet, plus la vraisamblance est proche de 1, plus le modèle est "proche" des données.
- La valeur d'une vraisemblance est difficile à interpréter (elle dépend notamment du nombre de données).
- La déviance permet de comparer la vraisemblance du modèle à celle d'un modèle parfait en terme d'adequation aux données : le modèle saturé.

La déviance

- Idée : se baser sur la vraisemblance. En effet, plus la vraisamblance est proche de 1, plus le modèle est "proche" des données.
- La valeur d'une vraisemblance est difficile à interpréter (elle dépend notamment du nombre de données).
- La déviance permet de comparer la vraisemblance du modèle à celle d'un modèle parfait en terme d'adequation aux données : le modèle saturé.

Le modèle saturé

- C'est le modèle qui ajuste "parfaitement" les observations. Il faut dissocier les types de données pour le définir.

Données individuelles

On note $(x_1, Y_1), \ldots, (x_n, Y_n)$ l'échantillon (tous les x_i sont différents). Le modèle saturé modélise la loi des Y_i par des Bernoulli de paramètre $p_{sat}(x_i)$ estimés selon $\hat{p}_{sat}(x_i) = Y_i$.

Données répétées

On note $(x_1, n_1, Y_1), \ldots, (x_T, n_T, Y_T)$ l'échantillon. Le modèle saturé modélise la loi des Y_t par des Binomiale de paramètres $(n_t, p_{sat}(x_t))$ avec $\hat{p}_{sat}(x_t) = Y_t/n_t$.

Le modèle saturé

- C'est le modèle qui ajuste "parfaitement" les observations. Il faut dissocier les types de données pour le définir.

Données individuelles

On note $(x_1, Y_1), \ldots, (x_n, Y_n)$ l'échantillon (tous les x_i sont différents). Le modèle saturé modélise la loi des Y_i par des Bernoulli de paramètre $p_{sat}(x_i)$ estimés selon $\hat{p}_{sat}(x_i) = Y_i$.

Données répétées

On note $(x_1, n_1, Y_1), \ldots, (x_T, n_T, Y_T)$ l'échantillon. Le modèle saturé modélise la loi des Y_t par des Binomiale de paramètres $(n_t, p_{sat}(x_t))$ avec $\hat{p}_{sat}(x_t) = Y_t/n_t$.

- On désigne par $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé calculé au point défini par les $\hat{p}_{sat}(x_i)$.

Propriété

- Dans le cas de données individuelles, $\mathcal{L}_{sat} = 0$.
- Pour des données répétées, on a

$$\mathcal{L}_{sat} = \sum_{t=1}^{T} \log \binom{n_t}{y_t} + \sum_{t=1}^{T} y_t \log \hat{p}_{sat}(x_t) + (n_t - y_t) \log(1 - \hat{p}_{sat}(x_t)).$$

Remarque

- En terme d'ajustement, on ne peut pas faire mieux que le modèle saturé.
- Néanmoins, ce modèle n'est généralement pas bon : il est sur-paramétré (il contient autant de paramètres que de points d'observations), d'où son nom.

- On désigne par $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé calculé au point défini par les $\hat{p}_{sat}(x_i)$.

Propriété

- Dans le cas de données individuelles, $\mathcal{L}_{sat} = 0$.
- Pour des données répétées, on a

$$\mathcal{L}_{sat} = \sum_{t=1}^{T} \log \binom{n_t}{y_t} + \sum_{t=1}^{T} y_t \log \hat{p}_{sat}(x_t) + (n_t - y_t) \log(1 - \hat{p}_{sat}(x_t)).$$

Remarque

- En terme d'ajustement, on ne peut pas faire mieux que le modèle saturé.
- Néanmoins, ce modèle n'est généralement pas bon : il est sur-paramétré (il contient autant de paramètres que de points d'observations), d'où son nom.

- On désigne par $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé calculé au point défini par les $\hat{p}_{sat}(x_i)$.

Propriété

- Dans le cas de données individuelles, $\mathcal{L}_{sat} = 0$.
- Pour des données répétées, on a

$$\mathcal{L}_{sat} = \sum_{t=1}^{T} \log \binom{n_t}{y_t} + \sum_{t=1}^{T} y_t \log \hat{p}_{sat}(x_t) + (n_t - y_t) \log(1 - \hat{p}_{sat}(x_t)).$$

Remarque

- En terme d'ajustement, on ne peut pas faire mieux que le modèle saturé.
- Néanmoins, ce modèle n'est généralement pas bon : il est sur-paramétré (il contient autant de paramètres que de points d'observations), d'où son nom.

- On désigne par $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé calculé au point défini par les $\hat{p}_{sat}(x_i)$.

Propriété

- Dans le cas de données individuelles, $\mathcal{L}_{sat} = 0$.
- Pour des données répétées, on a

$$\mathcal{L}_{sat} = \sum_{t=1}^{T} \log \binom{n_t}{y_t} + \sum_{t=1}^{T} y_t \log \hat{p}_{sat}(x_t) + (n_t - y_t) \log(1 - \hat{p}_{sat}(x_t)).$$

Remarque

- En terme d'ajustement, on ne peut pas faire mieux que le modèle saturé.
- Néanmoins, ce modèle n'est généralement pas bon : il est sur-paramétré (il contient autant de paramètres que de points d'observations), d'où son nom.

Déviance

- On note $\mathcal{M}$ un modèle logistique, $\hat{\beta}_n$ l'emv des paramètres et $\mathcal{L}_n$ la log-vraisemblance de ce modèle.

Définition

La déviance de $\mathcal{M}$ est définie par

$$D_{\mathcal{M}} = 2(\mathcal{L}_{sat} - \mathcal{L}_n(\hat{\beta}_n)).$$

- La déviance est positive $D_{\mathcal{M}} \geq 0$.
- Plus la déviance est faible, meilleur est le modèle en terme d'ajustement.

Déviance

- On note $\mathcal{M}$ un modèle logistique, $\hat{\beta}_n$ l'emv des paramètres et $\mathcal{L}_n$ la log-vraisemblance de ce modèle.

Définition

La déviance de $\mathcal{M}$ est définie par

$$D_{\mathcal{M}} = 2(\mathcal{L}_{sat} - \mathcal{L}_n(\hat{\beta}_n)).$$

- La déviance est positive $D_{\mathcal{M}} \geq 0$.
- Plus la déviance est faible, meilleur est le modèle en terme d'ajustement.

Illustration

- On reprend le jeu de données sur le "role des femmes" dans la société (données répétées).
- La déviance est présente dans les sorties de la fonction glm :

```
> model1 <- glm(cbind(agree,disagree)~sex+education,data=womensrole,family=binomial
> model1

Call:  glm(formula = cbind(agree, disagree) ~ sex + education,
         family = binomial, data = womensrole)

Coefficients:
(Intercept)    sexFemale    education
    2.74796     -0.04349     -0.28970

Degrees of Freedom: 29 Total (i.e. Null);  27 Residual
Null Deviance:      398.9
 Residual Deviance: 36.89   AIC: 165.4
```

- On peut également la récupérer avec la fonction deviance

```
> deviance(model1)
[1] 36.89419
```

Illustration

- On reprend le jeu de données sur le "role des femmes" dans la société (données répétées).
- La déviance est présente dans les sorties de la fonction glm :

```
> model1 <- glm(cbind(agree,disagree)~sex+education,data=womensrole,family=binomial
> model1

Call:  glm(formula = cbind(agree, disagree) ~ sex + education,
         family = binomial, data = womensrole)

Coefficients:
(Intercept)    sexFemale    education
    2.74796     -0.04349     -0.28970

Degrees of Freedom: 29 Total (i.e. Null);  27 Residual
Null Deviance:      398.9
 Residual Deviance: 36.89   AIC: 165.4
```

- On peut également la récupérer avec la fonction deviance

```
> deviance(model1)
[1] 36.89419
```

Le test d'adéquation de la déviance

- Idée : déviance faible $\Longrightarrow$ bonne adéquation.
- On pose H_0 : "le modèle est adéquat" (les données sont bien générées selon le modèle logistique en question) contre H_1 : "il ne l'est pas".

Propriété

En présence de données répétées, la déviance suit une loi du χ^2_{T-p} sous H_0 lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si $D_{\mathcal{M},obs}$ est plus grande que le quantile d'ordre $1-\alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> 1-pchisq(model1$deviance,model1$df.resid)
[1] 0.09705949
```

Le test d'adéquation de la déviance

- Idée : déviance faible $\Longrightarrow$ bonne adéquation.
- On pose H_0 : "le modèle est adéquat" (les données sont bien générées selon le modèle logistique en question) contre H_1 : "il ne l'est pas".

Propriété

En présence de données répétées, la déviance suit une loi du χ^2_{T-p} sous H_0 lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si $D_{\mathcal{M},obs}$ est plus grande que le quantile d'ordre $1-\alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> 1-pchisq(model1$deviance,model1$df.resid)
[1] 0.09705949
```

Le test d'adéquation de la déviance

- Idée : déviance faible $\Longrightarrow$ bonne adéquation.
- On pose H_0 : "le modèle est adéquat" (les données sont bien générées selon le modèle logistique en question) contre H_1 : "il ne l'est pas".

Propriété

En présence de données répétées, la déviance suit une loi du χ^2_{T-p} sous H_0 lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si $D_{\mathcal{M},obs}$ est plus grande que le quantile d'ordre $1-\alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> 1-pchisq(model1$deviance,model1$df.resid)
[1] 0.09705949
```

Le test d'adéquation de la déviance

- Idée : déviance faible $\Longrightarrow$ bonne adéquation.
- On pose H_0 : "le modèle est adéquat" (les données sont bien générées selon le modèle logistique en question) contre H_1 : "il ne l'est pas".

Propriété

En présence de données répétées, la déviance suit une loi du χ^2_{T-p} sous H_0 lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si $D_{\mathcal{M},obs}$ est plus grande que le quantile d'ordre $1-\alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> 1-pchisq(model1$deviance,model1$df.resid)
[1] 0.09705949
```

Le test d'adéquation de la déviance

- Idée : déviance faible $\Longrightarrow$ bonne adéquation.
- On pose H_0 : "le modèle est adéquat" (les données sont bien générées selon le modèle logistique en question) contre H_1 : "il ne l'est pas".

Propriété

En présence de données répétées, la déviance suit une loi du χ^2_{T-p} sous H_0 lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si $D_{\mathcal{M},obs}$ est plus grande que le quantile d'ordre $1-\alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> 1-pchisq(model1$deviance,model1$df.resid)
[1] 0.09705949
```

Le test d'adéquation de Pearson

- Permet de tester les mêmes hypothèses que précédemment et toujours pour des données répétées.
- La statistique de test est la suivante :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(x_t))^2}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}.$$

Propriété

En présence de données répétées, P suit une loi du χ^2_{T-p} sous H_0 en présence de données répétées lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si P_{obs} est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> P <- sum(residuals(model1,type="pearson")^2)
> 1-pchisq(P,nrow(womensrole)-length(model1$coef))
```

Le test d'adéquation de Pearson

- Permet de tester les mêmes hypothèses que précédemment et toujours pour des données répétées.
- La statistique de test est la suivante :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(x_t))^2}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}.$$

Propriété

En présence de données répétées, P suit une loi du χ^2_{T-p} sous H_0 en présence de données répétées lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si P_{obs} est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> P <- sum(residuals(model1,type="pearson")^2)
> 1-pchisq(P,nrow(womensrole)-length(model1$coef))
```

Le test d'adéquation de Pearson

- Permet de tester les mêmes hypothèses que précédemment et toujours pour des données répétées.
- La statistique de test est la suivante :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(x_t))^2}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}.$$

Propriété

En présence de données répétées, P suit une loi du χ^2_{T-p} sous H_0 en présence de données répétées lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si P_{obs} est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> P <- sum(residuals(model1,type="pearson")^2)
> 1-pchisq(P,nrow(womensrole)-length(model1$coef))
```

Le test d'adéquation de Pearson

- Permet de tester les mêmes hypothèses que précédemment et toujours pour des données répétées.
- La statistique de test est la suivante :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(x_t))^2}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}.$$

Propriété

En présence de données répétées, P suit une loi du χ^2_{T-p} sous H_0 en présence de données répétées lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si P_{obs} est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> P <- sum(residuals(model1,type="pearson")^2)
> 1-pchisq(P,nrow(womensrole)-length(model1$coef))
```

Le test d'adéquation de Pearson

- Permet de tester les mêmes hypothèses que précédemment et toujours pour des données répétées.
- La statistique de test est la suivante :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(x_t))^2}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}.$$

Propriété

En présence de données répétées, P suit une loi du χ^2_{T-p} sous H_0 en présence de données répétées lorsque $n_t \to \infty, t = 1, \ldots, T$.

- Conclusion : on rejette H_0 si P_{obs} est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p}.
- Sur R, on calcule la probabilité critique avec

```
> P <- sum(residuals(model1,type="pearson")^2)
> 1-pchisq(P,nrow(womensrole)-length(model1$coef))
```

Commentaires

- Les deux tests d'adéquation sont asymptotiques et utilisables uniquement dans le cas de données répétées.
- Il faut par conséquent avoir suffisamment d'observations en chaque points du design pour pouvoir les appliquer.
- Le test de déviance est généralement privilégié.
- En présence de données individuelles, on utilise souvent le test de Hosmer et Lemeshow : l'approche consiste à regrouper les données et à définir une statistique de test de type Pearson.

Commentaires

- Les deux tests d'adéquation sont asymptotiques et utilisables uniquement dans le cas de données répétées.
- Il faut par conséquent avoir suffisamment d'observations en chaque points du design pour pouvoir les appliquer.
- Le test de déviance est généralement privilégié.
- En présence de données individuelles, on utilise souvent le test de Hosmer et Lemeshow : l'approche consiste à regrouper les données et à définir une statistique de test de type Pearson.

Commentaires

- Les deux tests d'adéquation sont asymptotiques et utilisables uniquement dans le cas de données répétées.
- Il faut par conséquent avoir suffisamment d'observations en chaque points du design pour pouvoir les appliquer.
- Le test de déviance est généralement privilégié.
- En présence de données individuelles, on utilise souvent le test de Hosmer et Lemeshow : l'approche consiste à regrouper les données et à définir une statistique de test de type Pearson.

Commentaires

- Les deux tests d'adéquation sont asymptotiques et utilisables uniquement dans le cas de données répétées.
- Il faut par conséquent avoir suffisamment d'observations en chaque points du design pour pouvoir les appliquer.
- Le test de déviance est généralement privilégié.
- En présence de données individuelles, on utilise souvent le test de Hosmer et Lemeshow : l'approche consiste à regrouper les données et à définir une statistique de test de type Pearson.

Test d'Hosmer Lemeshow

On est en présence de données individuelles $(x_1, Y_1), \ldots, (x_n, Y_n)$. La statistique de test se construit comme suit.

1. Les probabilités estimées $p_{\hat{\beta}_n}(x_i)$ sont ordonnées par ordre croissant.
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
   - m_k les effectifs du groupe k ;
   - o_k le nombre de succès $(Y = 1)$ observé dans le groupe k ;
   - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k.

- La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k\mu_k)^2}{m_k\mu_k(1-\mu_k)}.$$

- Le test se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2_{K-2}.

Test d'Hosmer Lemeshow

On est en présence de données individuelles $(x_1, Y_1), \ldots, (x_n, Y_n)$. La statistique de test se construit comme suit.

1. Les probabilités estimées $p_{\hat{\beta}_n}(x_i)$ sont ordonnées par ordre croissant.
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
   - m_k les effectifs du groupe k ;
   - o_k le nombre de succès $(Y = 1)$ observé dans le groupe k ;
   - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k.

- La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k\mu_k)^2}{m_k\mu_k(1-\mu_k)}.$$

- Le test se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2_{K-2}.

Test d'Hosmer Lemeshow

On est en présence de données individuelles $(x_1, Y_1), \ldots, (x_n, Y_n)$. La statistique de test se construit comme suit.

1. Les probabilités estimées $p_{\hat{\beta}_n}(x_i)$ sont ordonnées par ordre croissant.
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
   - m_k les effectifs du groupe k ;
   - o_k le nombre de succès $(Y = 1)$ observé dans le groupe k ;
   - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k.

- La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k\mu_k)^2}{m_k\mu_k(1-\mu_k)}.$$

- Le test se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2_{K-2}.

Test d'Hosmer Lemeshow

On est en présence de données individuelles $(x_1, Y_1), \ldots, (x_n, Y_n)$. La statistique de test se construit comme suit.

1. Les probabilités estimées $p_{\hat{\beta}_n}(x_i)$ sont ordonnées par ordre croissant.
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
   - m_k les effectifs du groupe k ;
   - o_k le nombre de succès $(Y = 1)$ observé dans le groupe k ;
   - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k.

- La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k\mu_k)^2}{m_k\mu_k(1-\mu_k)}.$$

- Le test se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2_{K-2}.

Illustration sous R

- Sous R, on peut effectuer le test avec la fonction HLgof.test du package MKmisc.
- On teste le modèle sélectionné par la fonction bestglm sur l'exemple de la maladie cardiovasculaire :

```
> library(MKmisc)
> HLgof.test(fit=fitted(model4),obs=dapp$chd)
$C

Hosmer-Lemeshow C statistic

data:  fitted(model4) and dapp$chd
X-squared = 3.9695, df = 8, p-value = 0.8599


$H

Hosmer-Lemeshow H statistic

data:  fitted(model4) and dapp$chd
X-squared = 4.5285, df = 8, p-value = 0.8066
```

Illustration sous R

- Sous R, on peut effectuer le test avec la fonction HLgof.test du package MKmisc.
- On teste le modèle sélectionné par la fonction bestglm sur l'exemple de la maladie cardiovasculaire :

```
> library(MKmisc)
> HLgof.test(fit=fitted(model4),obs=dapp$chd)
$C

Hosmer-Lemeshow C statistic

data:  fitted(model4) and dapp$chd
X-squared = 3.9695, df = 8, p-value = 0.8599


$H

Hosmer-Lemeshow H statistic

data:  fitted(model4) and dapp$chd
X-squared = 4.5285, df = 8, p-value = 0.8066
```

Illustration sous R

- Sous R, on peut effectuer le test avec la fonction HLgof.test du package MKmisc.
- On teste le modèle sélectionné par la fonction bestglm sur l'exemple de la maladie cardiovasculaire :

```
> library(MKmisc)
> HLgof.test(fit=fitted(model4),obs=dapp$chd)
$C

Hosmer-Lemeshow C statistic

data:  fitted(model4) and dapp$chd
X-squared = 3.9695, df = 8, p-value = 0.8599


$H

Hosmer-Lemeshow H statistic

data:  fitted(model4) and dapp$chd
X-squared = 4.5285, df = 8, p-value = 0.8066
```

- L'analyse des résidus permet, dans une certaine mesure, d'affiner un modèle.
- Elle permet de détecter des individus atypiques ou aberrants ou encore de détecter des effets non linéaires.
- On distingue plusieurs types de résidus que nous présentons dans le cas de données répétées.

- L'analyse des résidus permet, dans une certaine mesure, d'affiner un modèle.
- Elle permet de détecter des individus atypiques ou aberrants ou encore de détecter des effets non linéaires.
- On distingue plusieurs types de résidus que nous présentons dans le cas de données répétées.

- L'analyse des résidus permet, dans une certaine mesure, d'affiner un modèle.
- Elle permet de détecter des individus atypiques ou aberrants ou encore de détecter des effets non linéaires.
- On distingue plusieurs types de résidus que nous présentons dans le cas de données répétées.

Les résidus de Pearson

On désigne par $(x_t, n_t, Y_t), t = 1, \ldots, T$ les données.

- Les résidus de Pearson sont définis par :

$$Rp_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}}.$$

- Lorsque n_t est grand, la loi de Rp_t est proche d'une $\mathcal{N}(0, 1)$. On peut ainsi analyser les résidus de Pearson de la même manière que les résidus du modèle linéaire Gaussien.
- La statistique de Pearson s'exprime en fonction des résidus de Pearson $P = \sum_{t=1}^{T} Rp_t^2$.

Les résidus de Pearson

On désigne par $(x_t, n_t, Y_t), t = 1, \ldots, T$ les données.

- Les résidus de Pearson sont définis par :

$$Rp_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}}.$$

- Lorsque n_t est grand, la loi de Rp_t est proche d'une $\mathcal{N}(0, 1)$. On peut ainsi analyser les résidus de Pearson de la même manière que les résidus du modèle linéaire Gaussien.
- La statistique de Pearson s'exprime en fonction des résidus de Pearson $P = \sum_{t=1}^{T} Rp_t^2$.

Les résidus de Pearson

On désigne par $(x_t, n_t, Y_t), t = 1, \ldots, T$ les données.

- Les résidus de Pearson sont définis par :

$$Rp_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))}}.$$

- Lorsque n_t est grand, la loi de Rp_t est proche d'une $\mathcal{N}(0, 1)$. On peut ainsi analyser les résidus de Pearson de la même manière que les résidus du modèle linéaire Gaussien.
- La statistique de Pearson s'exprime en fonction des résidus de Pearson $P = \sum_{t=1}^{T} Rp_t^2$.

Version standardisée

- Les résidus de Pearson définis précédemment ne sont pas de variance 1.
- Il est souvent préférable d'utiliser une version standardisée de ces résidus. Pour ce faire, on remarque que

$$\mathbf{V}[Y_t - np_{\hat{\beta}_n}(x_t)] \approx n_t p_\beta(x_t)(1 - p_\beta(x_t))(1 - h_t),$$

où h_t est le $t$eme élément de la diagonale de

$$\mathbb{H} = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}_n} \mathbb{X})^{-1} \mathbb{X}' W_{\hat{\beta}_n}.$$

Définition

Les résidus de Pearson standardisés sont définis par

$$Rps_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))(1 - h_t)}}, t = 1, \ldots, T.$$

Version standardisée

- Les résidus de Pearson définis précédemment ne sont pas de variance 1.
- Il est souvent préférable d'utiliser une version standardisée de ces résidus. Pour ce faire, on remarque que

$$\mathbf{V}[Y_t - np_{\hat{\beta}_n}(x_t)] \approx n_t p_\beta(x_t)(1 - p_\beta(x_t))(1 - h_t),$$

où h_t est le $t$eme élément de la diagonale de

$$\mathbb{H} = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}_n} \mathbb{X})^{-1} \mathbb{X}' W_{\hat{\beta}_n}.$$

Définition

Les résidus de Pearson standardisés sont définis par

$$Rps_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))(1 - h_t)}}, t = 1, \ldots, T.$$

Version standardisée

- Les résidus de Pearson définis précédemment ne sont pas de variance 1.
- Il est souvent préférable d'utiliser une version standardisée de ces résidus. Pour ce faire, on remarque que

$$\mathbf{V}[Y_t - np_{\hat{\beta}_n}(x_t)] \approx n_t p_\beta(x_t)(1 - p_\beta(x_t))(1 - h_t),$$

où h_t est le $t$eme élément de la diagonale de

$$\mathbb{H} = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}_n} \mathbb{X})^{-1} \mathbb{X}' W_{\hat{\beta}_n}.$$

Définition

Les résidus de Pearson standardisés sont définis par

$$Rps_t = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{\sqrt{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))(1 - h_t)}}, t = 1, \ldots, T.$$

Résidus de déviance

- Les résidus de déviance sont définis par

$$Rd_t = \sqrt{2\left[y_t \log \frac{\bar{Y}_t}{p_{\hat{\beta}_n}(x_t)} + (n_t - Y_t) \log \frac{n_t - Y_t}{n_t - n_t p_{\hat{\beta}_n}(x_t)}\right]}.$$

- Lorsque n_t est grand, la loi de Rd_t est proche d'une $\mathcal{N}(0,1)$. On peut ainsi analyser les résidus de déviance de la même manière que les résidus du modèle linéaire Gaussien.
- La déviance s'exprime en fonction des résidus de déviance $D = \sum_{t=1}^{T} Rd_t^2$.

- Là encore, il existe une version standardisée :

$$Rds_t = \frac{Rd_t}{\sqrt{1-h_t}}, \quad t = 1, \ldots, T.$$

Résidus de déviance

- Les résidus de déviance sont définis par

$$Rd_t = \sqrt{2\left[y_t \log \frac{\bar{Y}_t}{p_{\hat{\beta}_n}(x_t)} + (n_t - Y_t) \log \frac{n_t - Y_t}{n_t - n_t p_{\hat{\beta}_n}(x_t)}\right]}.$$

- Lorsque n_t est grand, la loi de Rd_t est proche d'une $\mathcal{N}(0,1)$. On peut ainsi analyser les résidus de déviance de la même manière que les résidus du modèle linéaire Gaussien.
- La déviance s'exprime en fonction des résidus de déviance $D = \sum_{t=1}^{T} Rd_t^2$.
- Là encore, il existe une version standardisée :

$$Rds_t = \frac{Rd_t}{\sqrt{1-h_t}}, \quad t = 1, \ldots, T.$$

Résidus de déviance

- Les résidus de déviance sont définis par

$$Rd_t = \sqrt{2\left[y_t \log \frac{\bar{Y}_t}{p_{\hat{\beta}_n}(x_t)} + (n_t - Y_t) \log \frac{n_t - Y_t}{n_t - n_t p_{\hat{\beta}_n}(x_t)}\right]}.$$

- Lorsque n_t est grand, la loi de Rd_t est proche d'une $\mathcal{N}(0,1)$. On peut ainsi analyser les résidus de déviance de la même manière que les résidus du modèle linéaire Gaussien.
- La déviance s'exprime en fonction des résidus de déviance $D = \sum_{t=1}^{T} Rd_t^2$.

- Là encore, il existe une version standardisée :

$$Rds_t = \frac{Rd_t}{\sqrt{1-h_t}}, \quad t = 1, \ldots, T.$$

Résidus de déviance

- Les résidus de déviance sont définis par

$$Rd_t = \sqrt{2\left[y_t \log \frac{\bar{Y}_t}{p_{\hat{\beta}_n}(x_t)} + (n_t - Y_t) \log \frac{n_t - Y_t}{n_t - n_t p_{\hat{\beta}_n}(x_t)}\right]}.$$

- Lorsque n_t est grand, la loi de Rd_t est proche d'une $\mathcal{N}(0,1)$. On peut ainsi analyser les résidus de déviance de la même manière que les résidus du modèle linéaire Gaussien.
- La déviance s'exprime en fonction des résidus de déviance $D = \sum_{t=1}^{T} Rd_t^2$.

- Là encore, il existe une version standardisée :

$$Rds_t = \frac{Rd_t}{\sqrt{1-h_t}}, \quad t = 1, \ldots, T.$$

Examen des résidus

- Les diagnostics sont essentiellement graphiques :
  1. **Index plot** : numéro de l'observation en abcisse, valeur du résidu en ordonnée.
  2. **Prédiction/résidus** : probabilité prédite au point x_t en abcisse et résidu en ordonnée.
- On pourra identifier :
  1. Les valeurs élevées de résidus (individus atypiques...)
  2. Structures sur le nuage des résidus (si c'est le cas il faudra envisager de modifier la combinaison linéaire des variables explicatives)

Remarque importante

- Les résultats énoncés sur les résidus (Pearson ou déviance) sont vraies lorsque n_t est grand...
- Dans le cas de données individuelles, on observera (quasi)-systématiquement des structurations sur les nuages de résidus.

Examen des résidus

- Les diagnostics sont essentiellement graphiques :
  1. **Index plot** : numéro de l'observation en abcisse, valeur du résidu en ordonnée.
  2. **Prédiction/résidus** : probabilité prédite au point x_t en abcisse et résidu en ordonnée.
- On pourra identifier :
  1. Les valeurs élevées de résidus (individus atypiques...)
  2. Structures sur le nuage des résidus (si c'est le cas il faudra envisager de modifier la combinaison linéaire des variables explicatives)

Remarque importante

- Les résultats énoncés sur les résidus (Pearson ou déviance) sont vraies lorsque n_t est grand...
- Dans le cas de données individuelles, on observera (quasi)-systématiquement des structurations sur les nuages de résidus.

Examen des résidus

- Les diagnostics sont essentiellement graphiques :
  1. **Index plot** : numéro de l'observation en abcisse, valeur du résidu en ordonnée.
  2. **Prédiction/résidus** : probabilité prédite au point x_t en abcisse et résidu en ordonnée.
- On pourra identifier :
  1. Les valeurs élevées de résidus (individus atypiques...)
  2. Structures sur le nuage des résidus (si c'est le cas il faudra envisager de modifier la combinaison linéaire des variables explicatives)

Remarque importante

- Les résultats énoncés sur les résidus (Pearson ou déviance) sont vraies lorsque n_t est grand...
- Dans le cas de données individuelles, on observera (quasi)-systématiquement des structurations sur les nuages de résidus.

Examen des résidus

- Les diagnostics sont essentiellement graphiques :
  1. **Index plot** : numéro de l'observation en abcisse, valeur du résidu en ordonnée.
  2. **Prédiction/résidus** : probabilité prédite au point x_t en abcisse et résidu en ordonnée.
- On pourra identifier :
  1. Les valeurs élevées de résidus (individus atypiques...)
  2. Structures sur le nuage des résidus (si c'est le cas il faudra envisager de modifier la combinaison linéaire des variables explicatives)

Remarque importante

- Les résultats énoncés sur les résidus (Pearson ou déviance) sont vraies lorsque n_t est grand...
- Dans le cas de données individuelles, on observera (quasi)-systématiquement des structurations sur les nuages de résidus.

Examen des résidus

- Les diagnostics sont essentiellement graphiques :
  1. **Index plot** : numéro de l'observation en abcisse, valeur du résidu en ordonnée.
  2. **Prédiction/résidus** : probabilité prédite au point x_t en abcisse et résidu en ordonnée.
- On pourra identifier :
  1. Les valeurs élevées de résidus (individus atypiques...)
  2. Structures sur le nuage des résidus (si c'est le cas il faudra envisager de modifier la combinaison linéaire des variables explicatives)

Remarque importante

- Les résultats énoncés sur les résidus (Pearson ou déviance) sont vraies lorsque n_t est grand...
- Dans le cas de données individuelles, on observera (quasi)-systématiquement des structurations sur les nuages de résidus.

Exemple

- On reprend les données womensrole et on considère le modèle logistique

```
> model1 <- glm(cbind(agree,disagree)~sex+education,data=womensrole,
                                                        family=binomial)
```

- Les fonctions residuals et rstandard permettent de calculer les différents type des résidus ainsi que leur version standardisée.

```
> res1 <- residuals(model1,type="deviance") #résidus de déviance
> res2 <- rstandard(model1,type="deviance") #résidus de déviance standardisés
```

- On trace les graphes avec

```
> par(mfrow=c(1,2))
> plot(res2,ylab="Residuals")
> abline(h=c(-2,2))
> plot(predict(model1,type="r"),res2,xlab="Fitted values",ylab="Residuals")
> abline(h=c(-2,2))
```

Exemple

- On reprend les données womensrole et on considère le modèle logistique

```
> model1 <- glm(cbind(agree,disagree)~sex+education,data=womensrole,
                                                    family=binomial)
```

- Les fonctions residuals et rstandard permettent de calculer les différents type des résidus ainsi que leur version standardisée.

```
> res1 <- residuals(model1,type="deviance") #résidus de déviance
> res2 <- rstandard(model1,type="deviance") #résidus de déviance standardisés
```

- On trace les graphes avec

```
> par(mfrow=c(1,2))
> plot(res2,ylab="Residuals")
> abline(h=c(-2,2))
> plot(predict(model1,type="r"),res2,xlab="Fitted values",ylab="Residuals")
> abline(h=c(-2,2))
```

Exemple

- On reprend les données womensrole et on considère le modèle logistique

```
> model1 <- glm(cbind(agree,disagree)~sex+education,data=womensrole,
                                                    family=binomial)
```

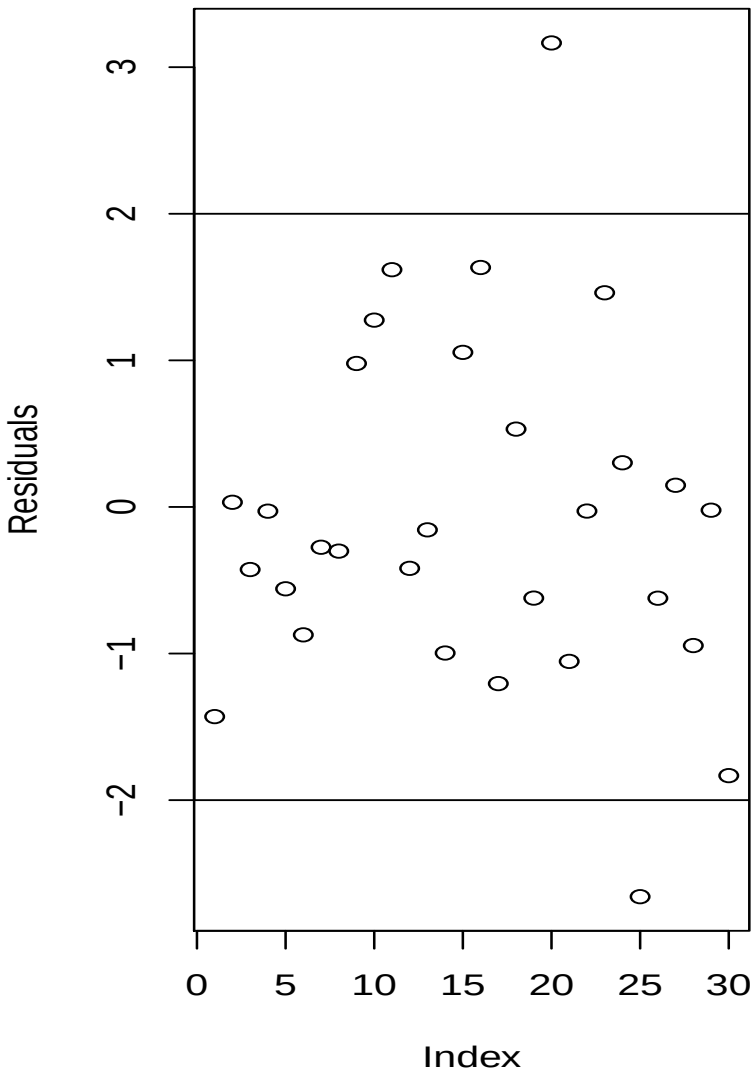
- Les fonctions residuals et rstandard permettent de calculer les différents type des résidus ainsi que leur version standardisée.

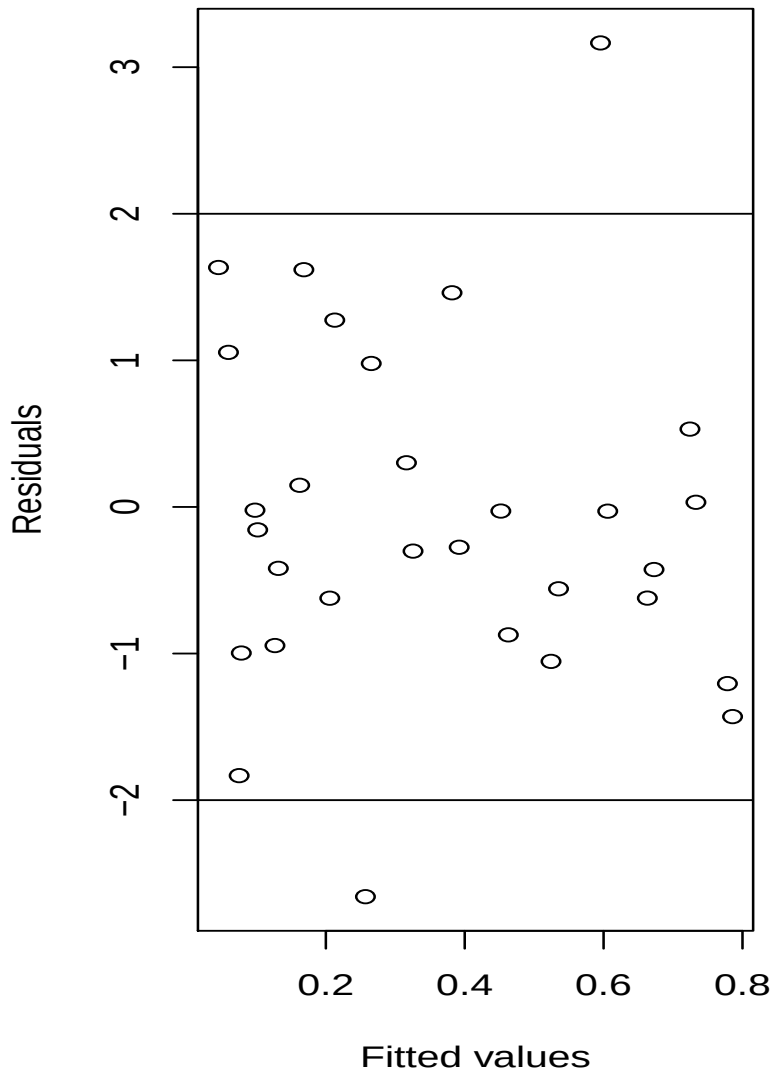
```
> res1 <- residuals(model1,type="deviance") #résidus de déviance
> res2 <- rstandard(model1,type="deviance") #résidus de déviance standardisés
```

- On trace les graphes avec

```
> par(mfrow=c(1,2))
> plot(res2,ylab="Residuals")
> abline(h=c(-2,2))
> plot(predict(model1,type="r"),res2,xlab="Fitted values",ylab="Residuals")
> abline(h=c(-2,2))
```

Tracé des résidus





Résidus Partiels

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots, \beta_p x_p.$$

- Les résidus partiels sont définis par :

$$r_{tj} = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))} + \hat{\beta}_j x_{ij}, \quad t = 1, \ldots T, j = 1 \ldots p.$$

Diagnostic

- L'analyse consiste à tracer pour toutes les variables j les T résidus $r_{tj}, t = 1, \ldots T$.
- Si le tracé est linéaire alors tout est "normal". Si par contre une tendance non linéaire se dégage, il faut remplacer la variable j par une fonction de celle ci donnant la même tendance que celle observée.

Résidus Partiels

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots, \beta_p x_p.$$

- Les résidus partiels sont définis par :

$$r_{tj} = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))} + \hat{\beta}_j x_{ij}, \quad t = 1, \ldots T, j = 1 \ldots p.$$

Diagnostic

- L'analyse consiste à tracer pour toutes les variables j les T résidus $r_{tj}, t = 1, \ldots T$.
- Si le tracé est linéaire alors tout est "normal". Si par contre une tendance non linéaire se dégage, il faut remplacer la variable j par une fonction de celle ci donnant la même tendance que celle observée.

Résidus Partiels

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots, \beta_p x_p.$$

- Les résidus partiels sont définis par :

$$r_{tj} = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))} + \hat{\beta}_j x_{ij}, \quad t = 1, \ldots T, j = 1 \ldots p.$$

Diagnostic

- L'analyse consiste à tracer pour toutes les variables j les T résidus $r_{tj}, t = 1, \ldots T$.
- Si le tracé est linéaire alors tout est "normal". Si par contre une tendance non linéaire se dégage, il faut remplacer la variable j par une fonction de celle ci donnant la même tendance que celle observée.

Résidus Partiels

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots, \beta_p x_p.$$

- Les résidus partiels sont définis par :

$$r_{tj} = \frac{Y_t - n_t p_{\hat{\beta}_n}(x_t)}{n_t p_{\hat{\beta}_n}(x_t)(1 - p_{\hat{\beta}_n}(x_t))} + \hat{\beta}_j x_{ij}, \quad t = 1, \ldots T, j = 1 \ldots p.$$

Diagnostic

- L'analyse consiste à tracer pour toutes les variables j les T résidus $r_{tj}, t = 1, \ldots T$.
- Si le tracé est linéaire alors tout est "normal". Si par contre une tendance non linéaire se dégage, il faut remplacer la variable j par une fonction de celle ci donnant la même tendance que celle observée.

Exemple

- On considère le modèle logistique permettant d'expliquer `etat` par `marque` et age pour les données panne.
- La fonction residuals permet de calculer les résidus partiels

```
> model <- glm(etat~.,data=panne,family=binomial)
> residpartiel <- residuals(model,type="partial")
```

- On trace les résidus partiels pour la variable age avec :

```
> plot(panne$age,residpartiel[,"age"],cex=0.5)
> est <- loess(residpartiel[,"age"]~panne$age)
> ordre <- order(panne$age)
> matlines(panne$age[ordre],predict(est)[ordre])
> abline(lsfit(panne$age,residpartiel[,"age"]),lty=2)
```

Exemple

- On considère le modèle logistique permettant d'expliquer `etat` par `marque` et age pour les données panne.
- La fonction residuals permet de calculer les résidus partiels

```
> model <- glm(etat~.,data=panne,family=binomial)
> residpartiel <- residuals(model,type="partial")
```

- On trace les résidus partiels pour la variable age avec :

```
> plot(panne$age,residpartiel[,"age"],cex=0.5)
> est <- loess(residpartiel[,"age"]~panne$age)
> ordre <- order(panne$age)
> matlines(panne$age[ordre],predict(est)[ordre])
> abline(lsfit(panne$age,residpartiel[,"age"]),lty=2)
```

Exemple

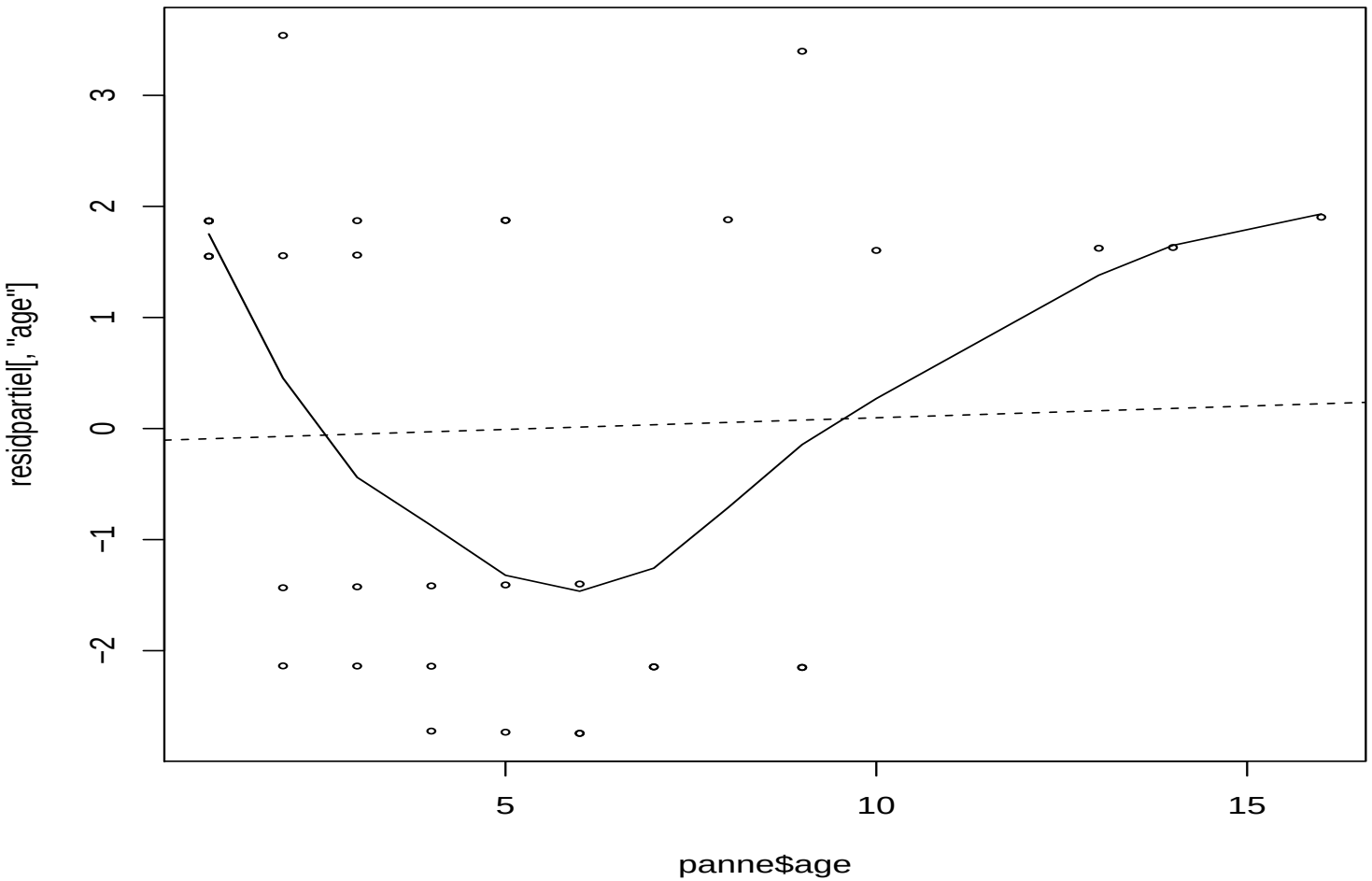
- On considère le modèle logistique permettant d'expliquer `etat` par `marque` et age pour les données panne.
- La fonction residuals permet de calculer les résidus partiels

```
> model <- glm(etat~.,data=panne,family=binomial)
> residpartiel <- residuals(model,type="partial")
```

- On trace les résidus partiels pour la variable age avec :

```
> plot(panne$age,residpartiel[,"age"],cex=0.5)
> est <- loess(residpartiel[,"age"]~panne$age)
> ordre <- order(panne$age)
> matlines(panne$age[ordre],predict(est)[ordre])
> abline(lsfit(panne$age,residpartiel[,"age"]),lty=2)
```

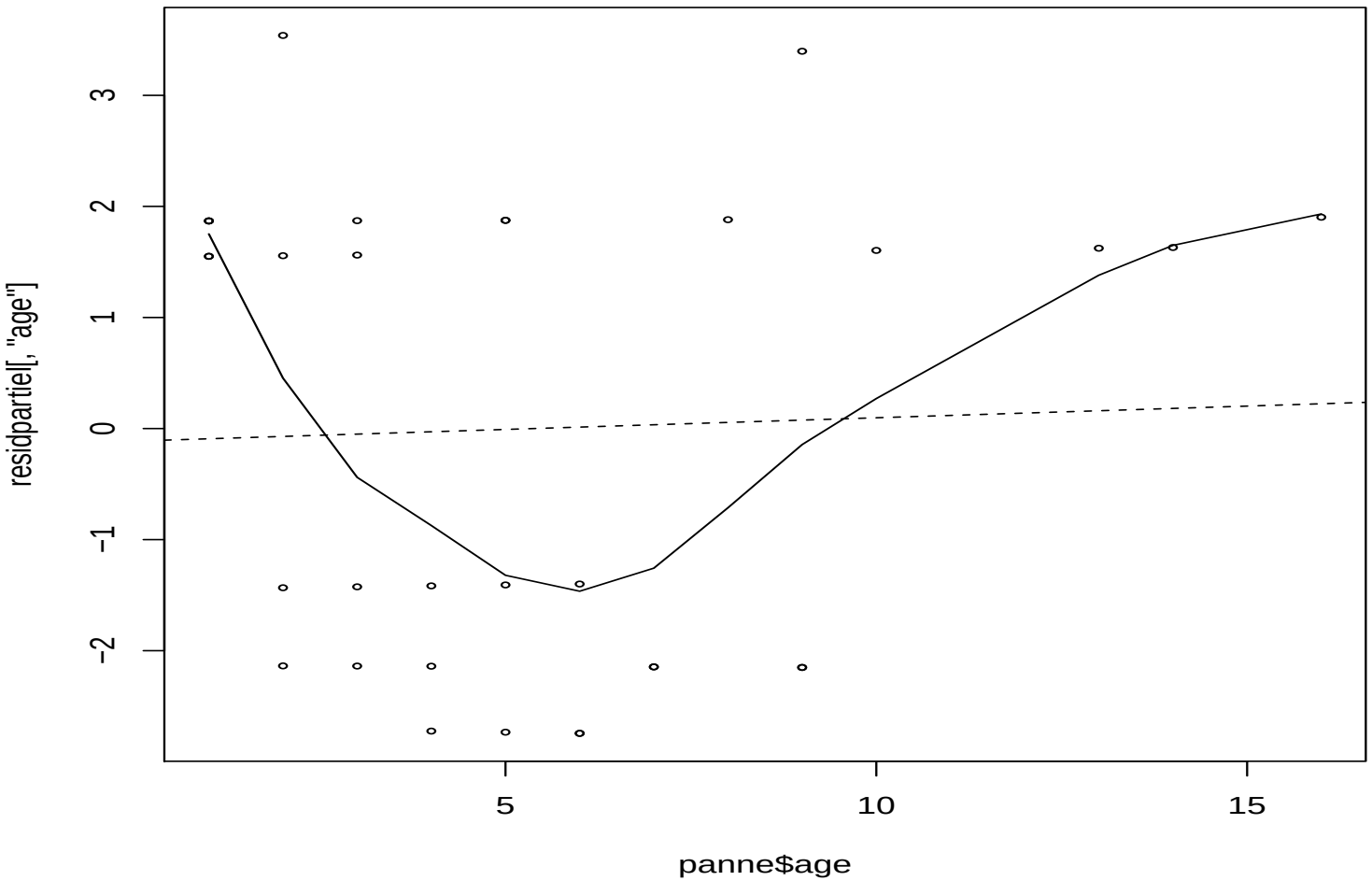
Tracé des résidus partiels



Conclusion

Le graphe suggère d'ajouter la variable age^2 dans le modèle.

Tracé des résidus partiels



Conclusion

Le graphe suggère d'ajouter la variable age^2 dans le modèle.

1. Quelques jeux de données

2. Sélection-choix de modèles
   - Critères de choix de modèles
     - Basés sur l'ajustement (AIC-BIC)
     - Basés sur la prévision (probabilité d'erreur)
   - Sélection de variables

3. Validation de modèles
   - Test d'adéquation de la déviance
   - Examen des résidus
   - Points leviers et points influents

Points leviers

- Ce sont les points du design qui déterminent fortement leur propre estimation.
- L'analyse est similaire à celle du modèle de régression linéaire.

- On rappelle que l'emv $\hat{\beta}_n$ s'écrit

$$\hat{\beta}_n = (\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z.$$

- La prédiction linéaire des individus est donc donnée par

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z = HZ,$$

- Et celle de l'individu i par

$$[\mathbb{X}\hat{\beta}_n]_i = H_{ii} Z_i + \sum_{j \neq i} H_{ij} Z_j.$$

Points leviers

- Ce sont les points du design qui déterminent fortement leur propre estimation.
- L'analyse est similaire à celle du modèle de régression linéaire.

- On rappelle que l'emv $\hat{\beta}_n$ s'écrit

$$\hat{\beta}_n = (\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z.$$

- La prédiction linéaire des individus est donc donnée par

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z = HZ,$$

- Et celle de l'individu i par

$$[\mathbb{X}\hat{\beta}_n]_i = H_{ii} Z_i + \sum_{j \neq i} H_{ij} Z_j.$$

Points leviers

- Ce sont les points du design qui déterminent fortement leur propre estimation.
- L'analyse est similaire à celle du modèle de régression linéaire.

- On rappelle que l'emv $\hat{\beta}_n$ s'écrit

$$\hat{\beta}_n = (\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z.$$

- La prédiction linéaire des individus est donc donnée par

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z = HZ,$$

- Et celle de l'individu i par

$$[\mathbb{X}\hat{\beta}_n]_i = H_{ii} Z_i + \sum_{j \neq i} H_{ij} Z_j.$$

Points leviers

- Ce sont les points du design qui déterminent fortement leur propre estimation.
- L'analyse est similaire à celle du modèle de régression linéaire.

- On rappelle que l'emv $\hat{\beta}_n$ s'écrit

$$\hat{\beta}_n = (\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z.$$

- La prédiction linéaire des individus est donc donnée par

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}' W_{\hat{\beta}} \mathbb{X})^{-1} \mathbb{X} W_{\hat{\beta}} Z = HZ,$$

- Et celle de l'individu i par

$$[\mathbb{X}\hat{\beta}_n]_i = H_{ii} Z_i + \sum_{j \neq i} H_{ij} Z_j.$$

- H étant un projecteur, on a $0 \leq H_{ii} \leq 1$. Par conséquent
  - Si $H_{ii} = 1$, alors $p_{\hat{\beta}_n}(x_i)$ est entirèrement déterminé par la $i$eme observation.
  - Si $H_{ii} = 0$, la $i$eme observation n'influence pas $p_{\hat{\beta}_n}(x_i)$.

Conclusion

- Pour mesurer l'influence d'une observation sur sa propre estimation, on représente le diagramme en batons des H_{ii}.
- On compare généralement la valeur des H_{ii} à $2p/n$ ou $3p/n$ pour déclarer les points comme leviers.

- H étant un projecteur, on a $0 \leq H_{ii} \leq 1$. Par conséquent
  - Si $H_{ii} = 1$, alors $p_{\hat{\beta}_n}(x_i)$ est entirèrement déterminé par la $i$eme observation.
  - Si $H_{ii} = 0$, la $i$eme observation n'influence pas $p_{\hat{\beta}_n}(x_i)$.

Conclusion

- Pour mesurer l'influence d'une observation sur sa propre estimation, on représente le diagramme en batons des H_{ii}.
- On compare généralement la valeur des H_{ii} à $2p/n$ ou $3p/n$ pour déclarer les points comme leviers.

- H étant un projecteur, on a $0 \leq H_{ii} \leq 1$. Par conséquent
  - Si $H_{ii} = 1$, alors $p_{\hat{\beta}_n}(x_i)$ est entirèrement déterminé par la $i$eme observation.
  - Si $H_{ii} = 0$, la $i$eme observation n'influence pas $p_{\hat{\beta}_n}(x_i)$.

Conclusion

- Pour mesurer l'influence d'une observation sur sa propre estimation, on représente le diagramme en batons des H_{ii}.
- On compare généralement la valeur des H_{ii} à $2p/n$ ou $3p/n$ pour déclarer les points comme leviers.

- H étant un projecteur, on a $0 \leq H_{ii} \leq 1$. Par conséquent
  - Si $H_{ii} = 1$, alors $p_{\hat{\beta}_n}(x_i)$ est entirèrement déterminé par la $i$eme observation.
  - Si $H_{ii} = 0$, la $i$eme observation n'influence pas $p_{\hat{\beta}_n}(x_i)$.

Conclusion

- Pour mesurer l'influence d'une observation sur sa propre estimation, on représente le diagramme en batons des H_{ii}.
- On compare généralement la valeur des H_{ii} à $2p/n$ ou $3p/n$ pour déclarer les points comme leviers.

- H étant un projecteur, on a $0 \leq H_{ii} \leq 1$. Par conséquent
  - Si $H_{ii} = 1$, alors $p_{\hat{\beta}_n}(x_i)$ est entirèrement déterminé par la $i$eme observation.
  - Si $H_{ii} = 0$, la $i$eme observation n'influence pas $p_{\hat{\beta}_n}(x_i)$.

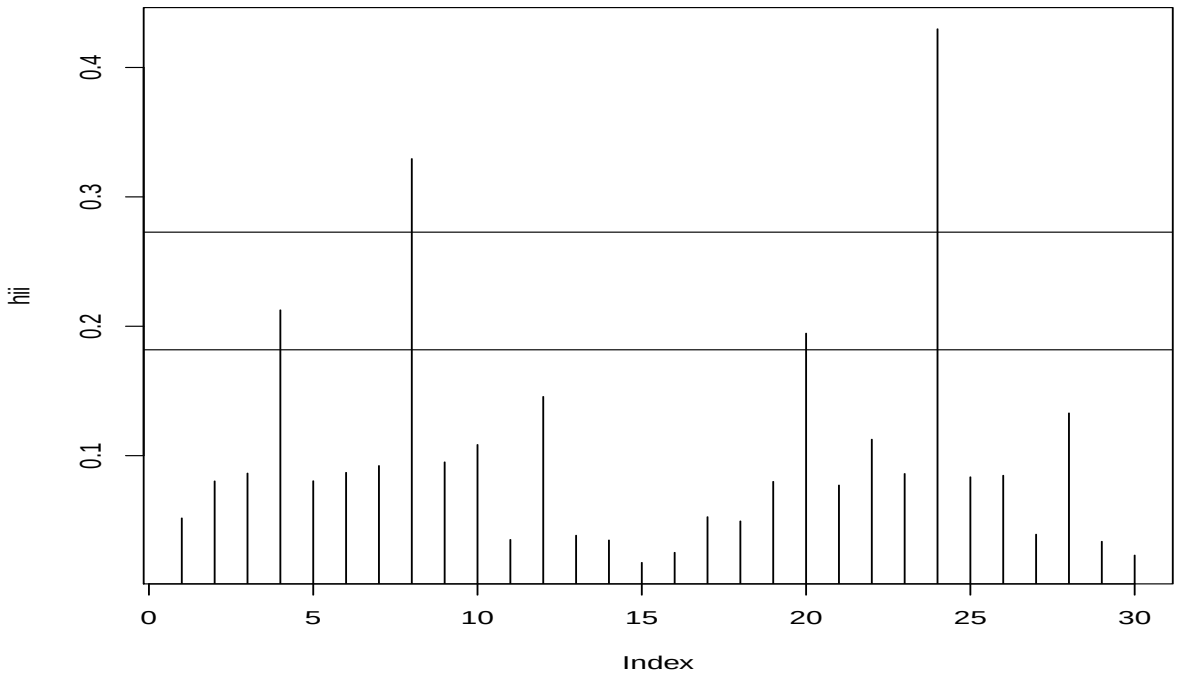
Conclusion

- Pour mesurer l'influence d'une observation sur sa propre estimation, on représente le diagramme en batons des H_{ii}.
- On compare généralement la valeur des H_{ii} à $2p/n$ ou $3p/n$ pour déclarer les points comme leviers.

Exemple

- On trace le diagramme en baton des H_{ii} pour le modèle construit sur les données womensrole.

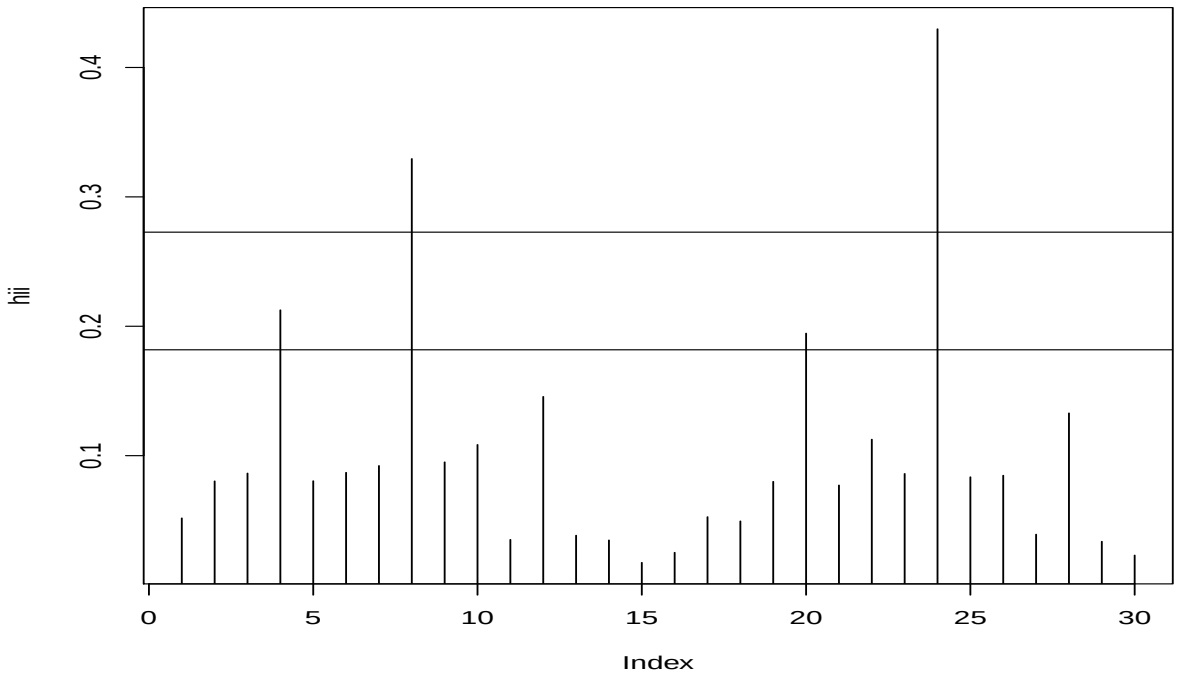
```
> model <- glm(cbind(agree,disagree)~sex+education,data=womensrole,
                                                    family=binomial)
> p <- length(model$coef)
> n <- nrow(panne)
> plot(influence(model)$hat,type="h",ylab="hii")
> abline(h=c(2*p/n,3*p/n))
```



Exemple

- On trace le diagramme en baton des H_{ii} pour le modèle construit sur les données womensrole.

```
> model <- glm(cbind(agree,disagree)~sex+education,data=womensrole,
                                                    family=binomial)
> p <- length(model$coef)
> n <- nrow(panne)
> plot(influence(model)$hat,type="h",ylab="hii")
> abline(h=c(2*p/n,3*p/n))
```



Points influents

- Les points influents sont des points qui influent sur le modèle de telle sorte que si on les enlève, alors l'estimation des coefficients sera fortement changée.
- La mesure la plus classique d'influence est la distance de Cook. Il s'agit d'une distance entre le coefficient estimé avec toutes les observations et celui estimé avec toutes les observations sauf une.

Définition

La distance de Cook pour l'individu i est définie par

$$DC_i = \frac{1}{p}(\hat{\beta}_{(i)} - \hat{\beta}_n)'\mathbb{X}'W_{\hat{\beta}}\mathbb{X}(\hat{\beta}_{(i)} - \hat{\beta}_n) \approx \frac{r_{Pi}^2 H_{ii}}{p(1-H_{ii})^2},$$

où r_{Pi} est le résidu de Pearson pour le $i$eme individu et $\hat{\beta}_{(i)}$ est l'emv calculé sans la $i$eme observation.

Points influents

- Les points influents sont des points qui influent sur le modèle de telle sorte que si on les enlève, alors l'estimation des coefficients sera fortement changée.
- La mesure la plus classique d'influence est la distance de Cook. Il s'agit d'une distance entre le coefficient estimé avec toutes les observations et celui estimé avec toutes les observations sauf une.

Définition

La distance de Cook pour l'individu i est définie par

$$DC_i = \frac{1}{p}(\hat{\beta}_{(i)} - \hat{\beta}_n)'\mathbb{X}'W_{\hat{\beta}}\mathbb{X}(\hat{\beta}_{(i)} - \hat{\beta}_n) \approx \frac{r_{Pi}^2 H_{ii}}{p(1-H_{ii})^2},$$

où r_{Pi} est le résidu de Pearson pour le $i$eme individu et $\hat{\beta}_{(i)}$ est l'emv calculé sans la $i$eme observation.

Points influents

- Les points influents sont des points qui influent sur le modèle de telle sorte que si on les enlève, alors l'estimation des coefficients sera fortement changée.
- La mesure la plus classique d'influence est la distance de Cook. Il s'agit d'une distance entre le coefficient estimé avec toutes les observations et celui estimé avec toutes les observations sauf une.

Définition

La distance de Cook pour l'individu i est définie par

$$DC_i = \frac{1}{p}(\hat{\beta}_{(i)} - \hat{\beta}_n)'\mathbb{X}'W_{\hat{\beta}}\mathbb{X}(\hat{\beta}_{(i)} - \hat{\beta}_n) \approx \frac{r_{Pi}^2 H_{ii}}{p(1-H_{ii})^2},$$

où r_{Pi} est le résidu de Pearson pour le $i$eme individu et $\hat{\beta}_{(i)}$ est l'emv calculé sans la $i$eme observation.

Distance de Cook

- Là encore, on représente la distance de Cook de chaque point du design à l'aide d'un diagramme en batons.

- Si une distance se révèle grande par rapport aux autres, alors ce point sera considéré comme influent. Il convient alors de comprendre pourquoi il est influent :
  - il est levier ;
  - il est aberrant ;
  - (les deux !)

Dans tous les cas il convient de comprendre si une erreur de mesure, une différence dans la population des individus est à l'origine de ce phénomène. Eventuellement pour obtenir des conclusions robustes il sera bon de refaire l'analyse sans ce(s) point(s).

Distance de Cook

- Là encore, on représente la distance de Cook de chaque point du design à l'aide d'un diagramme en batons.
- Si une distance se révèle grande par rapport aux autres, alors ce point sera considéré comme influent. Il convient alors de comprendre pourquoi il est influent :
  - il est levier ;
  - il est aberrant ;
  - (les deux !)

Dans tous les cas il convient de comprendre si une erreur de mesure, une différence dans la population des individus est à l'origine de ce phénomène. Eventuellement pour obtenir des conclusions robustes il sera bon de refaire l'analyse sans ce(s) point(s).

Distance de Cook

- Là encore, on représente la distance de Cook de chaque point du design à l'aide d'un diagramme en batons.

- Si une distance se révèle grande par rapport aux autres, alors ce point sera considéré comme influent. Il convient alors de comprendre pourquoi il est influent :
  - il est levier ;
  - il est aberrant ;
  - (les deux !)

Dans tous les cas il convient de comprendre si une erreur de mesure, une différence dans la population des individus est à l'origine de ce phénomène. Eventuellement pour obtenir des conclusions robustes il sera bon de refaire l'analyse sans ce(s) point(s).

Distance de Cook

- Là encore, on représente la distance de Cook de chaque point du design à l'aide d'un diagramme en batons.

- Si une distance se révèle grande par rapport aux autres, alors ce point sera considéré comme influent. Il convient alors de comprendre pourquoi il est influent :

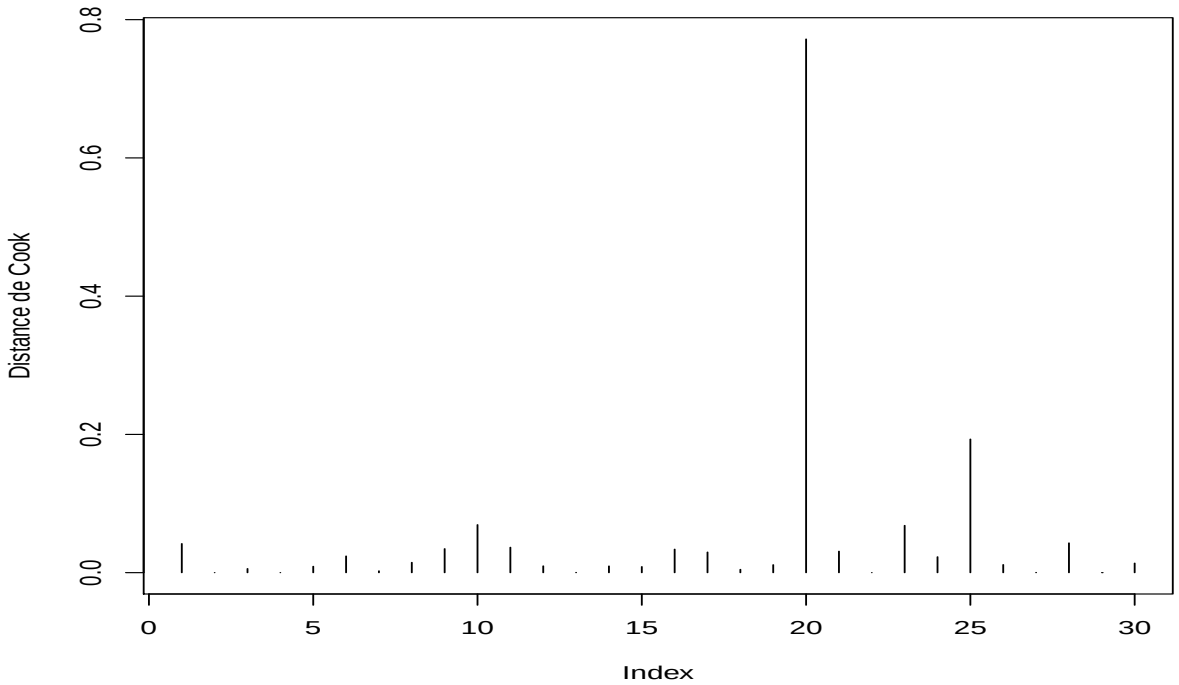
  - il est levier ;
  - il est aberrant ;
  - (les deux !)

Dans tous les cas il convient de comprendre si une erreur de mesure, une différence dans la population des individus est à l'origine de ce phénomène. Eventuellement pour obtenir des conclusions robustes il sera bon de refaire l'analyse sans ce(s) point(s).

Exemple

- La fonction cooks.distance permet de calculer les distances de Cook sur R :

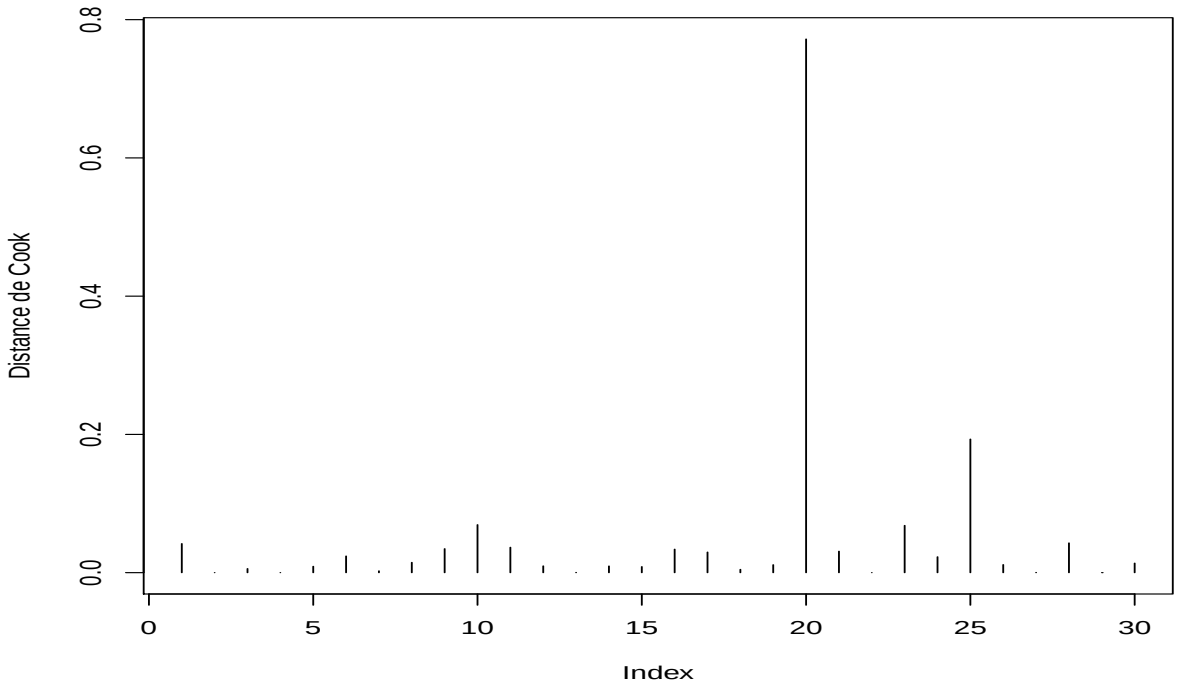
```
> plot(cooks.distance(model),type="h",ylab="Distance de Cook")
```



Exemple

- La fonction cooks.distance permet de calculer les distances de Cook sur R :

```
> plot(cooks.distance(model),type="h",ylab="Distance de Cook")
```



Quatrième partie IV

Quelques modèles logistiques polytomiques

Modèles polytomiques

- Jusqu'à présent, nous étions dans un contexte où la variable à expliquer était binaire à valeurs dans $\{0, 1\}$.
- Dans deux nombreux cas, on peut être amener à expliquer une variable qualitative prenant plus de deux modalités (scrutin à plus de deux candidats, degrés de satisfaction pour un produit, mention à un examen...)

Objectif

Etendre le modèle de régression logistique (binaire) à un cadre polytomique.

Modèles polytomiques

- Jusqu'à présent, nous étions dans un contexte où la variable à expliquer était binaire à valeurs dans $\{0, 1\}$.
- Dans deux nombreux cas, on peut être amener à expliquer une variable qualitative prenant plus de deux modalités (scrutin à plus de deux candidats, degrés de satisfaction pour un produit, mention à un examen...)

Objectif

Etendre le modèle de régression logistique (binaire) à un cadre polytomique.

Exemple

On pose à 195 étudiants la question : si vous trouvez un portefeuille dans la rue contenant de l'argent et des papiers :

- vous gardez tout (réponse 1) ;
- vous gardez l'argent et rendez le portefeuille (réponse 2) ;
- vous rendez tout (réponse 3).

On construit alors la variable `WALLET` telle que

- `WALLET`=1 si l'étudiant répond 1 ;
- `WALLET`=2 si l'étudiant répond 2 ;
- `WALLET`=3 si l'étudiant répond 3 ;

Exemple

On pose à 195 étudiants la question : si vous trouvez un portefeuille dans la rue contenant de l'argent et des papiers :

- vous gardez tout (réponse 1) ;
- vous gardez l'argent et rendez le portefeuille (réponse 2) ;
- vous rendez tout (réponse 3).

On construit alors la variable `WALLET` telle que

- `WALLET`=1 si l'étudiant répond 1 ;
- `WALLET`=2 si l'étudiant répond 2 ;
- `WALLET`=3 si l'étudiant répond 3 ;

Pour chaque étudiant, on note :

- Le sexe (variable `MALE`=1 si homme, 0 si femme) ;
- la nature des études suivies (variable `BUSINESS`= 1 pour les écoles de commerce, 0 pour les autres écoles) ;
- l'existence de punitions passées (variable `PUNISH`=1 si puni seulement à l'école primaire, 2 si puni seulement à l'école primaire et secondaire et 3 si puni seulement à l'école primaire, secondaire et supérieur) ;
- l'explication ou pas par les parents des punitions reçues dans l'enfance (variable `EXPLAIN`=1 si les parents expliquaient, 0 sinon).

On cherche à expliquer la variable `WALLET` par les autres variables.

Pour chaque étudiant, on note :

- Le sexe (variable `MALE`=1 si homme, 0 si femme) ;
- la nature des études suivies (variable `BUSINESS`= 1 pour les écoles de commerce, 0 pour les autres écoles) ;
- l'existence de punitions passées (variable `PUNISH`=1 si puni seulement à l'école primaire, 2 si puni seulement à l'école primaire et secondaire et 3 si puni seulement à l'école primaire, secondaire et supérieur) ;
- l'explication ou pas par les parents des punitions reçues dans l'enfance (variable `EXPLAIN`=1 si les parents expliquaient, 0 sinon).

On cherche à expliquer la variable `WALLET` par les autres variables.

Les données

- Les données brûtes sont présentées sous forme individuelles ($n = 195$) :

```
> donnees[1:5,]
  wallet male business punish explain
1      2    0        0      2       0
2      2    0        0      2       1
3      3    0        0      1       1
4      3    0        0      2       0
5      1    1        0      1       1
```

- Il y a cependant des répétitions... Et pour que certains indicateurs soient bien calculés (comme la déviance), il faut les présenter sous forde de données répétées ($T = 23$) :

```
> donnees1[1:5,]
  male business punish explain wallet.A A wallet.B B wallet.C  C
1    0        0      1       0        1 1        2 3        3  8
2    0        0      1       1        1 0        2 5        3 45
3    0        0      2       0        1 0        2 2        3  5
4    0        0      2       1        1 0        2 2        3  5
5    0        0      3       0        1 3        2 1        3  1
```

Les données

- Les données brûtes sont présentées sous forme individuelles ($n = 195$) :

```
> donnees[1:5,]
  wallet male business punish explain
1      2    0        0      2       0
2      2    0        0      2       1
3      3    0        0      1       1
4      3    0        0      2       0
5      1    1        0      1       1
```

- Il y a cependant des répétitions... Et pour que certains indicateurs soient bien calculés (comme la déviance), il faut les présenter sous forde de données répétées ($T = 23$) :

```
> donnees1[1:5,]
  male business punish explain wallet.A A wallet.B B wallet.C  C
1    0        0      1       0        1 1        2 3        3  8
2    0        0      1       1        1 0        2 5        3 45
3    0        0      2       0        1 0        2 2        3  5
4    0        0      2       1        1 0        2 2        3  5
5    0        0      3       0        1 3        2 1        3  1
```

Notations

- On cherche à expliquer une variable Y à K modalités $\{1, \ldots, K\}$ par p variables explicatives $X_1, \ldots, X_p$.
- Là encore, il convient de distinguer les données individuelles des données répétées.

- Données individuelles : il n'y a pas de répétitions sur les x_i. Les données sont $\{(x_1, y_1), \ldots, (x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \{1, \ldots, K\}\}$.
- Données répétées : $\{(x_t, n_t, s_{1t}, \ldots, s_{kt}), t = 1, \ldots, T\}$ où
  - n_t : nombre de répétitions au point x_t.
  - s_{jt} : nombre de fois où la modalité j de Y a été observée au point x_t :

$$s_{jt} = \sum_{i=1}^{n_t} \mathbf{1}_{y_{it}=j}, \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

Notations

- On cherche à expliquer une variable Y à K modalités $\{1, \ldots, K\}$ par p variables explicatives $X_1, \ldots, X_p$.
- Là encore, il convient de distinguer les données individuelles des données répétées.

- Données individuelles : il n'y a pas de répétitions sur les x_i. Les données sont $\{(x_1, y_1), \ldots, (x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \{1, \ldots, K\}\}$.
- Données répétées : $\{(x_t, n_t, s_{1t}, \ldots, s_{kt}), t = 1, \ldots, T\}$ où
  - n_t : nombre de répétitions au point x_t.
  - s_{jt} : nombre de fois où la modalité j de Y a été observée au point x_t :

$$s_{jt} = \sum_{i=1}^{n_t} \mathbf{1}_{y_{it}=j}, \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

Notations

- On cherche à expliquer une variable Y à K modalités $\{1, \ldots, K\}$ par p variables explicatives $X_1, \ldots, X_p$.
- Là encore, il convient de distinguer les données individuelles des données répétées.

- Données individuelles : il n'y a pas de répétitions sur les x_i. Les données sont $\{(x_1, y_1), \ldots, (x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \{1, \ldots, K\}\}$.
- Données répétées : $\{(x_t, n_t, s_{1t}, \ldots, s_{kt}), t = 1, \ldots, T\}$ où
  - n_t : nombre de répétitions au point x_t.
  - s_{jt} : nombre de fois où la modalité j de Y a été observée au point x_t :

$$s_{jt} = \sum_{i=1}^{n_t} \mathbf{1}_{y_{it}=j}, \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

Notations

- On cherche à expliquer une variable Y à K modalités $\{1, \ldots, K\}$ par p variables explicatives $X_1, \ldots, X_p$.
- Là encore, il convient de distinguer les données individuelles des données répétées.

- Données individuelles : il n'y a pas de répétitions sur les x_i. Les données sont $\{(x_1, y_1), \ldots, (x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \{1, \ldots, K\}\}$.
- Données répétées : $\{(x_t, n_t, s_{1t}, \ldots, s_{kt}), t = 1, \ldots, T\}$ où
  - n_t : nombre de répétitions au point x_t.
  - s_{jt} : nombre de fois où la modalité j de Y a été observée au point x_t :

$$s_{jt} = \sum_{i=1}^{n_t} \mathbf{1}_{y_{it}=j}, \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

Début de modélisation

- On se place dans le cas de données répétées : les s_{jt} sont des réalisations de variables aléatoires S_{jt}.
- Si les observations sont indépendantes alors le vecteur réponse $(S_{1t}, \ldots, S_{Kt})$ suit une loi multinomiale de paramètres $(n_t, p_1(x_t), \ldots, p_K(x_t))$ avec

$$p_j(x_t) = \mathbf{P}(Y_t = j).$$

Poser un modèle revient à spécifier des contraintes de forme entre les probabilités $p_j(x_t)$.

Début de modélisation

- On se place dans le cas de données répétées : les s_{jt} sont des réalisations de variables aléatoires S_{jt}.
- Si les observations sont indépendantes alors le vecteur réponse $(S_{1t}, \ldots, S_{Kt})$ suit une loi multinomiale de paramètres $(n_t, p_1(x_t), \ldots, p_K(x_t))$ avec

$$p_j(x_t) = \mathbf{P}(Y_t = j).$$

Poser un modèle revient à spécifier des contraintes de forme entre les probabilités $p_j(x_t)$.

Début de modélisation

- On se place dans le cas de données répétées : les s_{jt} sont des réalisations de variables aléatoires S_{jt}.
- Si les observations sont indépendantes alors le vecteur réponse $(S_{1t}, \ldots, S_{Kt})$ suit une loi multinomiale de paramètres $(n_t, p_1(x_t), \ldots, p_K(x_t))$ avec

$$p_j(x_t) = \mathbf{P}(Y_t = j).$$

Poser un modèle revient à spécifier des contraintes de forme entre les probabilités $p_j(x_t)$.

Les modèles étudiés

- Il y a plein de façons de poser des contraintes entre les probabilités $p_j(x_t)$.
- On peut donc définir un grand nombre de modèles.
- Nous présentons dans ce chapitre les modèles les plus classiques :

1. le modèle saturé ;
2. Le modèle de régression logistique multinomial ;
3. Le modèle logistique ordinal (ou modèle à égalité des pentes), valable dans le cas où Y est une variable ordinale.

Les modèles étudiés

- Il y a plein de façons de poser des contraintes entre les probabilités $p_j(x_t)$.
- On peut donc définir un grand nombre de modèles.
- Nous présontons dans ce chapitre les modèles les plus classiques :

1. le modèle saturé ;
2. Le modèle de régression logistique multinomial ;
3. Le modèle logistique ordinal (ou modèle à égalité des pentes), valable dans le cas où Y est une variable ordinale.

Présentation

- Tout comme dans le cas binaire, ce modèle ne pose pas de restrictions entre les probabilités $p_j(x_t)$.
- Il contient donc comme paramètres les probabilités

$$p_j(x_t) = \mathbf{P}(Y_t = j), \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

- Comme $\sum_{j=1}^{K} p_j(x_t) = 1$, ce modèle comprend $T(K-1)$ paramètres inconnus.

EMV

Il est facile de voir que les estimateurs du maximum de vraisemblance sont donnés par $\hat{p}_j(x_t) = S_{jt}/n_t$.

- Tout comme dans le cas binaire, ce modèle ne pose pas de restrictions entre les probabilités $p_j(x_t)$.
- Il contient donc comme paramètres les probabilités

$$p_j(x_t) = \mathbf{P}(Y_t = j), \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

- Comme $\sum_{j=1}^{K} p_j(x_t) = 1$, ce modèle comprend $T(K-1)$ paramètres inconnus.

EMV

Il est facile de voir que les estimateurs du maximum de vraisemblance sont donnés par $\hat{p}_j(x_t) = S_{jt}/n_t$.

- Tout comme dans le cas binaire, ce modèle ne pose pas de restrictions entre les probabilités $p_j(x_t)$.
- Il contient donc comme paramètres les probabilités

$$p_j(x_t) = \mathbf{P}(Y_t = j), \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

- Comme $\sum_{j=1}^{K} p_j(x_t) = 1$, ce modèle comprend $T(K-1)$ paramètres inconnus.

EMV

Il est facile de voir que les estimateurs du maximum de vraisemblance sont donnés par $\hat{p}_j(x_t) = S_{jt}/n_t$.

Présentation

- Tout comme dans le cas binaire, ce modèle ne pose pas de restrictions entre les probabilités $p_j(x_t)$.
- Il contient donc comme paramètres les probabilités

$$p_j(x_t) = \mathbf{P}(Y_t = j), \quad j = 1, \ldots, K, \quad t = 1, \ldots, T.$$

- Comme $\sum_{j=1}^{K} p_j(x_t) = 1$, ce modèle comprend $T(K-1)$ paramètres inconnus.

EMV

Il est facile de voir que les estimateurs du maximum de vraisemblance sont donnés par $\hat{p}_j(x_t) = S_{jt}/n_t$.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors

$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors

$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors

$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors
$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors
$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

Propriétés

- On note $\Pi_t = (p_1(x_t), \ldots, p_K(x_t))$ et $\hat{\Pi}_t = (\hat{p}_1(x_t), \ldots, \hat{p}_K(x_t))$.

Proposition

On a

1. $\hat{\Pi}_t$ est un estimateur sans biais de Π_t.
2. $\mathbf{V}(\hat{\Pi}_t) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(j,j) = p_j(x_t)(1-p_j(x_t))$ et $\Sigma_t(j,\ell) = -p_j(x_t)p_\ell(x_t)$ si $1 \leq j \neq \ell \leq K$.
3. Si $n_t \to \infty$ alors
$$\sqrt{n_t}(\hat{\Pi}_t - \Pi_t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Inconvénients

- On doit disposer d'un grand nombre de répétitions à chaque point du design pour que les estimateurs soient précis.
- Ce modèle ne nous renseigne pas sur les probabilités $\mathbf{P}(Y = j)$ pour des points x n'appartenant pas au design.

- On dispose d'une variable explicative Y à K modalités et on cherche à modéliser les probabilités

$$\mathbf{P}(Y_t = j), \quad j = 1, \ldots, K-1, \quad t = 1, \ldots, T.$$

- L'approche consiste à se donner une modalité de référence, par exemple la modalité K, et à modéliser les prbabilités $p_j(x)$ selon

$$\log \frac{p_j(x_t)}{p_K(x_t)} = \beta_{1j} x_{t1} + \ldots + \beta_{pj} x_{tp} = x_t' \beta_j,$$

où $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})$.

Remarques

- Le modèle ne dépend pas de la modalité de référence choisie ! (seule la valeur des coefficients, et donc leur interprétation, en dépend).
- Ce modèle comprend $p(K-1)$ paramètres à estimer.
- Si $K = 2$, on retombe sur le modèle logistique binaire.

- On dispose d'une variable explicative Y à K modalités et on cherche à modéliser les probabilités

$$\mathbf{P}(Y_t = j), \quad j = 1, \ldots, K-1, \quad t = 1, \ldots, T.$$

- L'approche consiste à se donner une modalité de référence, par exemple la modalité K, et à modéliser les prbabilités $p_j(x)$ selon

$$\log \frac{p_j(x_t)}{p_K(x_t)} = \beta_{1j} x_{t1} + \ldots + \beta_{pj} x_{tp} = x_t' \beta_j,$$

où $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})$.

Remarques

- Le modèle ne dépend pas de la modalité de référence choisie ! (seule la valeur des coefficients, et donc leur interprétation, en dépend).
- Ce modèle comprend $p(K-1)$ paramètres à estimer.
- Si $K = 2$, on retombe sur le modèle logistique binaire.

- On dispose d'une variable explicative Y à K modalités et on cherche à modéliser les probabilités

$$\mathbf{P}(Y_t = j), \quad j = 1, \ldots, K-1, \quad t = 1, \ldots, T.$$

- L'approche consiste à se donner une modalité de référence, par exemple la modalité K, et à modéliser les prbabilités $p_j(x)$ selon

$$\log \frac{p_j(x_t)}{p_K(x_t)} = \beta_{1j} x_{t1} + \ldots + \beta_{pj} x_{tp} = x_t' \beta_j,$$

où $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})$.

Remarques

- Le modèle ne dépend pas de la modalité de référence choisie ! (seule la valeur des coefficients, et donc leur interprétation, en dépend).
- Ce modèle comprend $p(K-1)$ paramètres à estimer.
- Si $K = 2$, on retombe sur le modèle logistique binaire.

- On dispose d'une variable explicative Y à K modalités et on cherche à modéliser les probabilités

$$\mathbf{P}(Y_t = j), \quad j = 1, \ldots, K-1, \quad t = 1, \ldots, T.$$

- L'approche consiste à se donner une modalité de référence, par exemple la modalité K, et à modéliser les prbabilités $p_j(x)$ selon

$$\log \frac{p_j(x_t)}{p_K(x_t)} = \beta_{1j} x_{t1} + \ldots + \beta_{pj} x_{tp} = x_t' \beta_j,$$

où $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})$.

Remarques

- Le modèle ne dépend pas de la modalité de référence choisie ! (seule la valeur des coefficients, et donc leur interprétation, en dépend).
- Ce modèle comprend $p(K-1)$ paramètres à estimer.
- Si $K = 2$, on retombe sur le modèle logistique binaire.

- On dispose d'une variable explicative Y à K modalités et on cherche à modéliser les probabilités

$$\mathbf{P}(Y_t = j), \quad j = 1, \ldots, K-1, \quad t = 1, \ldots, T.$$

- L'approche consiste à se donner une modalité de référence, par exemple la modalité K, et à modéliser les prbabilités $p_j(x)$ selon

$$\log \frac{p_j(x_t)}{p_K(x_t)} = \beta_{1j} x_{t1} + \ldots + \beta_{pj} x_{tp} = x_t' \beta_j,$$

où $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})$.

Remarques

- Le modèle ne dépend pas de la modalité de référence choisie ! (seule la valeur des coefficients, et donc leur interprétation, en dépend).
- Ce modèle comprend $p(K-1)$ paramètres à estimer.
- Si $K = 2$, on retombe sur le modèle logistique binaire.

Estimation des paramètres

- Tout se passe comme dans le cas binaire... mais les choses (vraisemblance, Information de Fisher...) sont beaucoup plus lourdes à écrire.
- Les paramètres inconnus du modèle sont estimés par maximum de vraisemblance et on montre que, sous des hypothèses similaires au cas binaire,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

- On en déduit des procédures de tests (Wald, rapport de vraisemblance, score) ainsi que des intervalles de confiance pour les paramètres du modèle.

Estimation des paramètres

- Tout se passe comme dans le cas binaire... mais les choses (vraisemblance, Information de Fisher...) sont beaucoup plus lourdes à écrire.
- Les paramètres inconnus du modèle sont estimés par maximum de vraisemblance et on montre que, sous des hypothèses similaires au cas binaire,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

- On en déduit des procédures de tests (Wald, rapport de vraisemblance, score) ainsi que des intervalles de confiance pour les paramètres du modèle.

Estimation des paramètres

- Tout se passe comme dans le cas binaire... mais les choses (vraisemblance, Information de Fisher...) sont beaucoup plus lourdes à écrire.
- Les paramètres inconnus du modèle sont estimés par maximum de vraisemblance et on montre que, sous des hypothèses similaires au cas binaire,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}).$$

- On en déduit des procédures de tests (Wald, rapport de vraisemblance, score) ainsi que des intervalles de confiance pour les paramètres du modèle.

Exemple

- Sur R, les fonctions multinom du package nnet ou vglm du package vgam permettent d'ajuster le modèle de régression logistique multinomial. Sous SAS, on pourra utiliser la proc CATMOD.

```
> model <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,multinomial
> model1 <- multinom(wallet~business+punish+male+explain,data=donnees)
> model
Call:
vglm(formula = cbind(A, B, C) ~ business + punish + male + explain,
    family = multinomial, data = donnees1)

Coefficients:
(Intercept):1 (Intercept):2   business1:1   business1:2     punish2:1
   -2.4062097    -1.1068174     1.1791118     0.4155709     1.1450946
    punish2:2     punish3:1     punish3:2       male1:1       male1:2
    0.2491692     2.1411710     0.3531734     1.2672026     1.1716184
   explain1:1    explain1:2
   -1.5935358    -0.7978215

Degrees of Freedom: 46 Total; 34 Residual
Residual deviance: 31.91969
Log-likelihood: -46.32335
```

- On peut tester le modèle contre la constante ou l'effet de chaque variable à l'aide d'un test de rapport de vraisemblance :

```
> lrtest_vglm(model)
Likelihood ratio test

Model 1: cbind(A, B, C) ~ business + punish + male + explain
Model 2: cbind(A, B, C) ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  34 -46.323
2  44 -71.613 10 50.579  2.088e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Anova(model1,type=3)
Analysis of Deviance Table (Type III tests)

Response: wallet
          LR Chisq Df Pr(>Chisq)
business   4.6540  2   0.097590 .
punish    11.0788  4   0.025692 *
male      13.0036  2   0.001501 **
explain    9.9181  2   0.007019 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Prévision

- Pour une nouvelle valeur de $x_{n+1} \in \mathbb{R}^p$, on peut naturellement estimer les probabilités que Y soit égales à j

$$p_j(x_{n+1}) = \mathbf{P}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\beta_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\beta_j)}$$

par

$$\hat{p}_j(x_{n+1}) = \hat{\mathbf{P}}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\hat{\beta}_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\hat{\beta}_j)}.$$

- Et connaissant la loi (asymptotique) des $\hat{\beta}_j$, on peut en déduire des intervalles de confiance pour $p_j(x_{n+1})$.

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Prévision

- Pour une nouvelle valeur de $x_{n+1} \in \mathbb{R}^p$, on peut naturellement estimer les probabilités que Y soit égales à j

$$p_j(x_{n+1}) = \mathbf{P}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\beta_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\beta_j)}$$

par

$$\hat{p}_j(x_{n+1}) = \hat{\mathbf{P}}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\hat{\beta}_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\hat{\beta}_j)}.$$

- Et connaissant la loi (asymptotique) des $\hat{\beta}_j$, on peut en déduire des intervalles de confiance pour $p_j(x_{n+1})$.

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Prévision

- Pour une nouvelle valeur de $x_{n+1} \in \mathbb{R}^p$, on peut naturellement estimer les probabilités que Y soit égales à j

$$p_j(x_{n+1}) = \mathbf{P}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\beta_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\beta_j)}$$

par

$$\hat{p}_j(x_{n+1}) = \hat{\mathbf{P}}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\hat{\beta}_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\hat{\beta}_j)}.$$

- Et connaissant la loi (asymptotique) des $\hat{\beta}_j$, on peut en déduire des intervalles de confiance pour $p_j(x_{n+1})$.

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Prévision

- Pour une nouvelle valeur de $x_{n+1} \in \mathbb{R}^p$, on peut naturellement estimer les probabilités que Y soit égales à j

$$p_j(x_{n+1}) = \mathbf{P}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\beta_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\beta_j)}$$

par

$$\hat{p}_j(x_{n+1}) = \hat{\mathbf{P}}(Y_{n+1} = j) = \frac{\exp(x'_{n+1}\hat{\beta}_j)}{1 + \sum_{j=1}^{K-1} \exp(x'_{n+1}\hat{\beta}_j)}.$$

- Et connaissant la loi (asymptotique) des $\hat{\beta}_j$, on peut en déduire des intervalles de confiance pour $p_j(x_{n+1})$.

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Rappel sur le modèle logistique binaire

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

Rappel sur le modèle logistique binaire

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

Rappel sur le modèle logistique binaire

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

On peut prolonger cette idée pour une variable Y ordinale à plus de deux modalités.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

alors le modèle est le modèle logistique.

- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

On peut prolonger cette idée pour une variable Y ordinale à plus de deux modalités.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

On peut prolonger cette idée pour une variable Y ordinale à plus de deux modalités.

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

  alors le modèle est le modèle logistique.
- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

On peut prolonger cette idée pour une variable Y ordinale à plus de deux modalités.

- On cherche à expliquer Y ordinale à valeurs dans $\{1, \ldots, K\}$ par p variables $X_1, \ldots, X_p$. On dispose d'un n-échantillon indépendant $(x_1, Y_1), \ldots, (x_n, Y_n)$.
- On suppose que Y_i est liée à une variable latente $Y_i^\star$ telle que

$$Y_i^\star = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

selon

$$Y_i = \begin{cases} 1 & \text{si } Y_i^\star < \alpha_1 \\ j & \text{si } \alpha_{j-1} \leq Y_i^\star < \alpha_j,\ j = 2, \ldots, K-1 \\ k & \text{si } Y_i^\star \geq \alpha_{K-1} \end{cases}$$

où $\alpha_1, \ldots, \alpha_{K-1}$ sont des seuils inconnus et ε est un terme d'erreur aléatoire.

Le modèle logistique ordinal

Si les ε sont de loi logistique, on obtient alors le modèle logistique ordinal

$$\text{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} = \alpha_j - x_i'\beta,\ j = 1, \ldots, K-1.$$

- On cherche à expliquer Y ordinale à valeurs dans $\{1, \ldots, K\}$ par p variables $X_1, \ldots, X_p$. On dispose d'un n-échantillon indépendant $(x_1, Y_1), \ldots, (x_n, Y_n)$.
- On suppose que Y_i est liée à une variable latente $Y_i^\star$ telle que

$$Y_i^\star = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

selon

$$Y_i = \begin{cases} 1 & \text{si } Y_i^\star < \alpha_1 \\ j & \text{si } \alpha_{j-1} \leq Y_i^\star < \alpha_j,\ j = 2, \ldots, K-1 \\ k & \text{si } Y_i^\star \geq \alpha_{K-1} \end{cases}$$

où $\alpha_1, \ldots, \alpha_{K-1}$ sont des seuils inconnus et ε est un terme d'erreur aléatoire.

Le modèle logistique ordinal

Si les ε sont de loi logistique, on obtient alors le modèle logistique ordinal

$$\text{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} = \alpha_j - x_i'\beta,\ j = 1, \ldots, K-1.$$

Remarques

- Ce modèle est aussi appelé modèle cumulatif (car il prend en compte les logit des probabilités $\{Y \leq j\}$ ou encore modèle logistique à égalité des pentes (on verra plus tard pourquoi).
- Les coefficients β associés aux variables explicatives ne dépendent pas de j, seules les constantes α_j en dépendent.
- Le modèle nécessite l'estimation de $p + K - 1$ paramètres.
- La paramétrisation peut varier selon les logiciels, la `proc logistic` de `SAS` ajuste par exemple le modèle

$$\mathrm{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \alpha_j + x_i'\beta,\ j = 1, \ldots, K-1.$$

Remarques

- Ce modèle est aussi appelé modèle cumulatif (car il prend en compte les logit des probabilités $\{Y \leq j\}$ ou encore modèle logistique à égalité des pentes (on verra plus tard pourquoi).
- Les coefficients β associés aux variables explicatives ne dépendent pas de j, seules les constantes α_j en dépendent.
- Le modèle nécessite l'estimation de $p + K - 1$ paramètres.
- La paramétrisation peut varier selon les logiciels, la `proc logistic` de `SAS` ajuste par exemple le modèle

$$\mathrm{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \alpha_j + x_i'\beta,\ j = 1, \ldots, K-1.$$

Remarques

- Ce modèle est aussi appelé modèle cumulatif (car il prend en compte les logit des probabilités $\{Y \leq j\}$ ou encore modèle logistique à égalité des pentes (on verra plus tard pourquoi).
- Les coefficients β associés aux variables explicatives ne dépendent pas de j, seules les constantes α_j en dépendent.
- Le modèle nécessite l'estimation de $p + K - 1$ paramètres.
- La paramétrisation peut varier selon les logiciels, la `proc logistic` de `SAS` ajuste par exemple le modèle

$$\mathrm{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \alpha_j + x_i'\beta,\ j = 1, \ldots, K-1.$$

Remarques

- Ce modèle est aussi appelé modèle cumulatif (car il prend en compte les logit des probabilités $\{Y \leq j\}$ ou encore modèle logistique à égalité des pentes (on verra plus tard pourquoi).
- Les coefficients β associés aux variables explicatives ne dépendent pas de j, seules les constantes α_j en dépendent.
- Le modèle nécessite l'estimation de $p + K - 1$ paramètres.
- La paramétrisation peut varier selon les logiciels, la `proc logistic` de `SAS` ajuste par exemple le modèle

$$\mathrm{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \alpha_j + x_i'\beta,\ j = 1, \ldots, K-1.$$

Exemples

- Sous R, on peut utiliser la fonction polr du package MASS

```
> model <- polr(wallet~.,data=donnees)
> model
Coefficients:
     male1  business1    punish2    punish3   explain1
-1.0598227 -0.7388746 -0.6276423 -1.4030892  1.0518775
Intercepts:
       1|2        2|3
-2.5678520 -0.7890143
```

- ou la fonction vglm du package VGAM

```
> model1 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                                    cumulative(parallel=TRUE))

> model1
Coefficients:
(Intercept):1 (Intercept):2     business1       punish2       punish3
   -2.5678316    -0.7889979     0.7388749     0.6276413     1.4030657
        male1      explain1
    1.0598180    -1.0518680
```

Moralité

Là encore, il convient d'aller voir dans l'aide la paramétrisation des fonctions.

Exemples

- Sous R, on peut utiliser la fonction polr du package MASS

```
> model <- polr(wallet~.,data=donnees)
> model
Coefficients:
     male1  business1    punish2    punish3   explain1
-1.0598227 -0.7388746 -0.6276423 -1.4030892  1.0518775
Intercepts:
       1|2        2|3
-2.5678520 -0.7890143
```

- ou la fonction vglm du package VGAM

```
> model1 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                                    cumulative(parallel=TRUE))

> model1
Coefficients:
(Intercept):1 (Intercept):2     business1       punish2       punish3
   -2.5678316    -0.7889979     0.7388749     0.6276413     1.4030657
        male1      explain1
    1.0598180    -1.0518680
```

Moralité

Là encore, il convient d'aller voir dans l'aide la paramétrisation des fonctions.

Exemples

- Sous R, on peut utiliser la fonction polr du package MASS

```
> model <- polr(wallet~.,data=donnees)
> model
Coefficients:
     male1  business1    punish2    punish3   explain1
-1.0598227 -0.7388746 -0.6276423 -1.4030892  1.0518775
Intercepts:
       1|2        2|3
-2.5678520 -0.7890143
```

- ou la fonction vglm du package VGAM

```
> model1 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                                    cumulative(parallel=TRUE))

> model1
Coefficients:
(Intercept):1 (Intercept):2     business1       punish2       punish3
   -2.5678316    -0.7889979     0.7388749     0.6276413     1.4030657
        male1      explain1
    1.0598180    -1.0518680
```

Moralité

Là encore, il convient d'aller voir dans l'aide la paramétrisation des fonctions.

Estimation des paramètres

- Les paramètres du modèle sont toujours estimés par maximum de vraisemblance et, sous des hypothèses similaires au cas binaire, on obtient la normalité asymptotique des estimateurs (on en déduit des IC et des procédures de test).
- Prévision : pour un nouvel individu x_{n+1} on pourra estimer les probabilités

$$\mathbf{P}_\beta(Y_{n+1} \leq j) = \frac{\exp(\alpha_j - x'_{n+1}\beta)}{1 + \exp(\alpha_j - x'_{n+1}\beta)}$$

par

$$\mathbf{P}_{\hat{\beta}}(Y_{n+1} \leq j) = \frac{\exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}{1 + \exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}$$

et en déduire ainsi les estimations des probabilités $\mathbf{P}_\beta(Y_{n+1} = j)$ pour faire la prévision.

Estimation des paramètres

- Les paramètres du modèle sont toujours estimés par maximum de vraisemblance et, sous des hypothèses similaires au cas binaire, on obtient la normalité asymptotique des estimateurs (on en déduit des IC et des procédures de test).
- Prévision : pour un nouvel individu x_{n+1} on pourra estimer les probabilités

$$\mathbf{P}_\beta(Y_{n+1} \leq j) = \frac{\exp(\alpha_j - x'_{n+1}\beta)}{1 + \exp(\alpha_j - x'_{n+1}\beta)}$$

par

$$\mathbf{P}_{\hat{\beta}}(Y_{n+1} \leq j) = \frac{\exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}{1 + \exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}$$

et en déduire ainsi les estimations des probabilités $\mathbf{P}_\beta(Y_{n+1} = j)$ pour faire la prévision.

Estimation des paramètres

- Les paramètres du modèle sont toujours estimés par maximum de vraisemblance et, sous des hypothèses similaires au cas binaire, on obtient la normalité asymptotique des estimateurs (on en déduit des IC et des procédures de test).
- Prévision : pour un nouvel individu x_{n+1} on pourra estimer les probabilités

$$\mathbf{P}_\beta(Y_{n+1} \leq j) = \frac{\exp(\alpha_j - x'_{n+1}\beta)}{1 + \exp(\alpha_j - x'_{n+1}\beta)}$$

par

$$\mathbf{P}_{\hat{\beta}}(Y_{n+1} \leq j) = \frac{\exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}{1 + \exp(\hat{\alpha}_j - x'_{n+1}\hat{\beta})}$$

et en déduire ainsi les estimations des probabilités $\mathbf{P}_\beta(Y_{n+1} = j)$ pour faire la prévision.

Tests de rapport de vraisemblance

- Comme pour le modèle logistique multinomial, on peut tester le modèle contre la constante ou l'effet de chaque variable à l'aide de tests de rapport de vraisemblance :

```
> lrtest_vglm(model1)
Likelihood ratio test

Model 1: cbind(A, B, C) ~ business + punish + male + explain
Model 2: cbind(A, B, C) ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  39 -49.211
2  44 -71.613  5 44.805   1.59e-08 ***

> Anova(model,type=3)
Analysis of Deviance Table (Type III tests)

Response: wallet
          LR Chisq Df Pr(>Chisq)
male       10.9265  1   0.000948 ***
business    4.2667  1   0.038867 *
punish      9.1512  2   0.010300 *
explain     9.5168  1   0.002036 **
```

Proportionnalité des odd ratio

- Pour $x_i \in \mathbb{R}^p$, on définit

$$odd(x_i; Y \leq j \text{ vs } Y > j) = \frac{\mathbf{P}_\beta(Y_i \leq j)}{\mathbf{P}_\beta(Y_i > j)}.$$

- On a alors que l'odd ratio

$$OR(x_i, x_k; Y \leq j \text{ vs } Y > j) = \exp((x_k - x_i)'\beta),$$

ne dépend pas de j. Cette propriété est appélée proportionnalité des odd ratio.

- Les logiciels renvoient souvent les OR associés à ces évènements (qui sont rarement simples à interpréter ! Voir TP2).

Proportionnalité des odd ratio

- Pour $x_i \in \mathbb{R}^p$, on définit

$$odd(x_i; Y \leq j \text{ vs } Y > j) = \frac{\mathbf{P}_\beta(Y_i \leq j)}{\mathbf{P}_\beta(Y_i > j)}.$$

- On a alors que l'odd ratio

$$OR(x_i, x_k; Y \leq j \text{ vs } Y > j) = \exp((x_k - x_i)'\beta),$$

ne dépend pas de j. Cette propriété est appélée proportionnalité des odd ratio.

- Les logiciels renvoient souvent les OR associés à ces évènements (qui sont rarement simples à interpréter ! Voir TP2).

Proportionnalité des odd ratio

- Pour $x_i \in \mathbb{R}^p$, on définit

$$odd(x_i; Y \leq j \text{ vs } Y > j) = \frac{\mathbf{P}_\beta(Y_i \leq j)}{\mathbf{P}_\beta(Y_i > j)}.$$

- On a alors que l'odd ratio

$$OR(x_i, x_k; Y \leq j \text{ vs } Y > j) = \exp((x_k - x_i)'\beta),$$

ne dépend pas de j. Cette propriété est appélée proportionnalité des odd ratio.

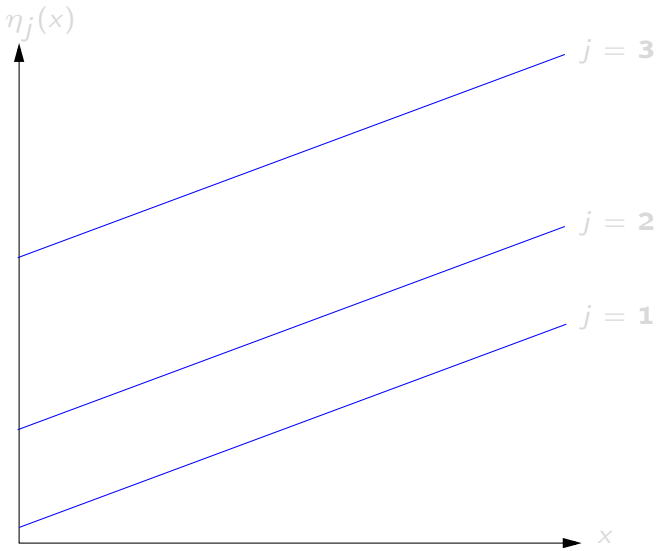
- Les logiciels renvoient souvent les OR associés à ces évènements (qui sont rarement simples à interpréter ! Voir TP2).

L'hypothèse d'égalité des pentes

- Supposons pour simplifier que l'on dispose d'une seule variable explicative X et considérons le modèle logistique ordinal

$$\operatorname{logit} \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta x_i.$$

- Seule la constante diffère selon j, c'est pourquoi on parle d'égalité des pentes.

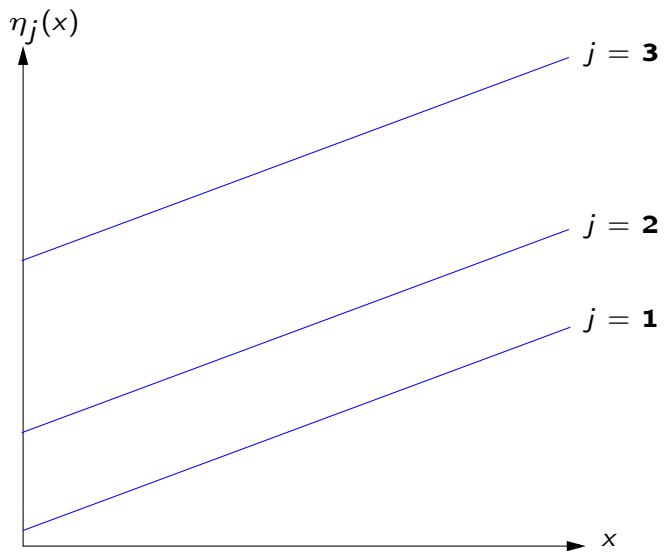


L'hypothèse d'égalité des pentes

- Supposons pour simplifier que l'on dispose d'une seule variable explicative X et considérons le modèle logistique ordinal

$$\operatorname{logit} \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta x_i.$$

- Seule la constante diffère selon j, c'est pourquoi on parle d'égalité des pentes.



- Si on lève cette propriété d'égalité des pentes (ce qui revient à considérer des pentes $\beta_j, j = 1, \ldots, K = 1$ différentes dans le modèle), on peut obtenir pour certaines valeurs de x_i

$$\mathbf{P}_\beta(Y_i \leq 1) > \mathbf{P}_\beta(Y_i \leq 2).$$

- Ce qui remet en cause le caractère ordinal de Y.
- Il est ainsi intéressant de développer un test permettant de vérifier l'égalité des pentes. On peut envisager des hypothèses du genre H_0 : "les pentes sont égales" contre H_1 : "elles ne le sont pas".

- Si on lève cette propriété d'égalité des pentes (ce qui revient à considérer des pentes $\beta_j, j = 1, \ldots, K = 1$ différentes dans le modèle), on peut obtenir pour certaines valeurs de x_i

$$\mathbf{P}_\beta(Y_i \leq 1) > \mathbf{P}_\beta(Y_i \leq 2).$$

- Ce qui remet en cause le caractère ordinal de Y.
- Il est ainsi intéressant de développer un test permettant de vérifier l'égalité des pentes. On peut envisager des hypothèses du genre H_0 : "les pentes sont égales" contre H_1 : "elles ne le sont pas".

- Si on lève cette propriété d'égalité des pentes (ce qui revient à considérer des pentes $\beta_j, j = 1, \ldots, K = 1$ différentes dans le modèle), on peut obtenir pour certaines valeurs de x_i

$$\mathbf{P}_\beta(Y_i \leq 1) > \mathbf{P}_\beta(Y_i \leq 2).$$

- Ce qui remet en cause le caractère ordinal de Y.
- Il est ainsi intéressant de développer un test permettant de vérifier l'égalité des pentes. On peut envisager des hypothèses du genre H_0 : "les pentes sont égales" contre H_1 : "elles ne le sont pas".

Le test d'égalité des pentes

- Il suffit de voir que le modèle logistique ordinal (modèle à égalité de pentes) est emboité dans le modèle

$$\text{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta_{1,j} x_{i1} - \ldots - \beta_{p,j} x_{ip}.$$

- Tester l'égalité des pentes revient donc à tester

$$H_0 : \beta_{\ell,1} = \ldots = \beta_{\ell,K-1}, \quad \forall \ell = 1, \ldots, p.$$

- On peut utiliser les statistiques de Wald, du rapport de vraisemblance ou du score pour effectuer ce test entre modèles emboités qui (pour n assez grand) suivent une loi χ^2 à

$$(K-1)(p-1) - (K-1+p) = p(K-2)$$

degrés de liberté.

Le test d'égalité des pentes

- Il suffit de voir que le modèle logistique ordinal (modèle à égalité de pentes) est emboité dans le modèle

$$\text{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta_{1,j} x_{i1} - \ldots - \beta_{p,j} x_{ip}.$$

- Tester l'égalité des pentes revient donc à tester

$$H_0 : \beta_{\ell,1} = \ldots = \beta_{\ell,K-1}, \quad \forall \ell = 1, \ldots, p.$$

- On peut utiliser les statistiques de Wald, du rapport de vraisemblance ou du score pour effectuer ce test entre modèles emboités qui (pour n assez grand) suivent une loi χ^2 à

$$(K-1)(p-1) - (K-1+p) = p(K-2)$$

degrés de liberté.

Le test d'égalité des pentes

- Il suffit de voir que le modèle logistique ordinal (modèle à égalité de pentes) est emboité dans le modèle

$$\text{logit}\, \mathbf{P}_\beta(Y_i \leq j) = \alpha_j - \beta_{1,j} x_{i1} - \ldots - \beta_{p,j} x_{ip}.$$

- Tester l'égalité des pentes revient donc à tester

$$H_0 : \beta_{\ell,1} = \ldots = \beta_{\ell,K-1}, \quad \forall \ell = 1, \ldots, p.$$

- On peut utiliser les statistiques de Wald, du rapport de vraisemblance ou du score pour effectuer ce test entre modèles emboités qui (pour n assez grand) suivent une loi χ^2 à

$$(K-1)(p-1) - (K-1+p) = p(K-2)$$

degrés de liberté.

Exemple

- On teste l'égalité des pentes pour le jeu de données `portefeuille` :

```
> model2 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                          cumulative(parallel=FALSE))
> statRV <- -2*(logLik(model1)-logLik(model2))
> 1-pchisq(statRV,df=length(coef(model2))-length(coef(model1)))
[1] 0.4053199
```

On accepte l'hypothèse d'égalité des pentes au seuil $\alpha = 5\%$.

Remarque

Sous SAS, la proc logistic utilise par défaut la statistique du score pour tester l'égalité des pentes (voir TP2).

Exemple

- On teste l'égalité des pentes pour le jeu de données `portefeuille` :

```
> model2 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                          cumulative(parallel=FALSE))
> statRV <- -2*(logLik(model1)-logLik(model2))
> 1-pchisq(statRV,df=length(coef(model2))-length(coef(model1)))
[1] 0.4053199
```

On accepte l'hypothèse d'égalité des pentes au seuil $\alpha = 5\%$.

Remarque

Sous SAS, la proc logistic utilise par défaut la statistique du score pour tester l'égalité des pentes (voir TP2).

Exemple

- On teste l'égalité des pentes pour le jeu de données `portefeuille` :

```
> model2 <- vglm(cbind(A,B,C)~business+punish+male+explain,data=donnees1,
                                          cumulative(parallel=FALSE))
> statRV <- -2*(logLik(model1)-logLik(model2))
> 1-pchisq(statRV,df=length(coef(model2))-length(coef(model1)))
[1] 0.4053199
```

On accepte l'hypothèse d'égalité des pentes au seuil $\alpha = 5\%$.

Remarque

Sous `SAS`, la proc logistic utilise par défaut la statistique du score pour tester l'égalité des pentes (voir TP2).

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Exemple

En présence de données répétées, sous le modèle logistique ordinal, la déviance suit (pour n_t assez grand), une loi du χ_2 à $T(K-1)-(p+K-1)$ ddl.

- Exemple : Tests d'adéquation de déviance et de Pearson pour le `model1`.

```
> deviance(model1) #statistique de la déviance
[1] 37.69427
> 1-pchisq(deviance(model1),df=39)
[1] 0.5293916
> model1@res.ss #statistique de Pearson
[1] 31.41399
> 1-pchisq(model1@res.ss,df=39)
[1] 0.8009696
```

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Exemple

En présence de données répétées, sous le modèle logistique ordinal, la déviance suit (pour n_t assez grand), une loi du χ_2 à $T(K-1)-(p+K-1)$ ddl.

- Exemple : Tests d'adéquation de déviance et de Pearson pour le `model1`.

```
> deviance(model1) #statistique de la déviance
[1] 37.69427
> 1-pchisq(deviance(model1),df=39)
[1] 0.5293916
> model1@res.ss #statistique de Pearson
[1] 31.41399
> 1-pchisq(model1@res.ss,df=39)
[1] 0.8009696
```

Remarque

Les notions de tests d'adéquation, de résidus, de critères de performance (AIC, BIC), de sélection de variables vues pour le cas binaire se généralisent "aisément" au modèle de régression logistique multinomial.

Exemple

En présence de données répétées, sous le modèle logistique ordinal, la déviance suit (pour n_t assez grand), une loi du χ_2 à $T(K-1)-(p+K-1)$ ddl.

- Exemple : Tests d'adéquation de déviance et de Pearson pour le `model1`.

```
> deviance(model1) #statistique de la déviance
[1] 37.69427
> 1-pchisq(deviance(model1),df=39)
[1] 0.5293916
> model1@res.ss #statistique de Pearson
[1] 31.41399
> 1-pchisq(model1@res.ss,df=39)
[1] 0.8009696
```

Cinquième partie V

Schéma d'échantillonnage rétrospectif

Exemple

- On cherche à expliquer le développement d'une maladie rare en fonction de l'hérédité.
- On réalise une étude auprès de $n = 500$ patients et on considère les variables :
  - `maladie` qui vaut pre si l'individu développe la maladie, abs sinon.
  - `here` qui vaut oui si un des parents a développé la maladie, non sinon.

```
> donnees
  here abs pre
1  oui  48 208
2  non 202  42
```

Exemple

- On cherche à expliquer le développement d'une maladie rare en fonction de l'hérédité.
- On réalise une étude auprès de $n = 500$ patients et on considère les variables :
  - `maladie` qui vaut pre si l'individu développe la maladie, abs sinon.
  - `here` qui vaut oui si un des parents a développé la maladie, non sinon.

```
> donnees
  here abs pre
1  oui  48 208
2  non 202  42
```

Exemple

- On cherche à expliquer le développement d'une maladie rare en fonction de l'hérédité.
- On réalise une étude auprès de $n = 500$ patients et on considère les variables :
  - `maladie` qui vaut pre si l'individu développe la maladie, abs sinon.
  - `here` qui vaut oui si un des parents a développé la maladie, non sinon.

```
> donnees
  here abs pre
1  oui  48 208
2  non 202  42
```

Un modèle logistique

- On explique `maladie` par `here` à l'aide d'un modèle de régression logistique.

```
> model <- glm(cbind(pre,abs)~.,data=donnees,family=binomial)
> model

Call:  glm(formula = cbind(pre, abs) ~ ., family = binomial, data = donnees)

Coefficients:
(Intercept)      hereoui
     -1.571        3.037

Degrees of Freedom: 1 Total (i.e. Null);  0 Residual
Null Deviance:      222
Residual Deviance: -2.62e-14  AIC: 14.9
```

Remarque

Le modèle logistique est saturé.

Un modèle logistique

- On explique `maladie` par `here` à l'aide d'un modèle de régression logistique.

```
> model <- glm(cbind(pre,abs)~.,data=donnees,family=binomial)
> model

Call:  glm(formula = cbind(pre, abs) ~ ., family = binomial, data = donnees)

Coefficients:
(Intercept)      hereoui
     -1.571        3.037

Degrees of Freedom: 1 Total (i.e. Null);  0 Residual
Null Deviance:      222
Residual Deviance: -2.62e-14  AIC: 14.9
```

Remarque

Le modèle logistique est saturé.

- On calcule la probabilité prédite par le modèle d'être atteint en fonction de l'hérédité :

```
> predict(model,type="response")
        1         2
0.8125000 0.1721311
```

Commentaire

La maladie étant rare, les probabilités d'être atteint paraissent anormalement élevées...

Cause

- L'échantillon ne parait pas être représentatif de la population.
- Les personnes atteintes sont en effet sur-représentées.
- C'est souvent le cas dans les études cas-témoin, on essaie de rééquilibrer les proportions de personnes atteintes et non atteintes dans l'échantillon.

- On calcule la probabilité prédite par le modèle d'être atteint en fonction de l'hérédité :

```
> predict(model,type="response")
        1         2
0.8125000 0.1721311
```

Commentaire

La maladie étant rare, les probabilités d'être atteint paraissent anormalement élevées...

Cause

- L'échantillon ne parait pas être représentatif de la population.
- Les personnes atteintes sont en effet sur-représentées.
- C'est souvent le cas dans les études cas-témoin, on essaie de rééquilibrer les proportions de personnes atteintes et non atteintes dans l'échantillon.

- On calcule la probabilité prédite par le modèle d'être atteint en fonction de l'hérédité :

```
> predict(model,type="response")
        1         2
0.8125000 0.1721311
```

Commentaire

La maladie étant rare, les probabilités d'être atteint paraissent anormalement élevées...

Cause

- L'échantillon ne parait pas être représentatif de la population.
- Les personnes atteintes sont en effet sur-représentées.
- C'est souvent le cas dans les études cas-témoin, on essaie de rééquilibrer les proportions de personnes atteintes et non atteintes dans l'échantillon.

- On calcule la probabilité prédite par le modèle d'être atteint en fonction de l'hérédité :

```
> predict(model,type="response")
        1         2
0.8125000 0.1721311
```

Commentaire

La maladie étant rare, les probabilités d'être atteint paraissent anormalement élevées...

Cause

- L'échantillon ne parait pas être représentatif de la population.
- Les personnes atteintes sont en effet sur-représentées.
- C'est souvent le cas dans les études cas-témoin, on essaie de rééquilibrer les proportions de personnes atteintes et non atteintes dans l'échantillon.

- On calcule la probabilité prédite par le modèle d'être atteint en fonction de l'hérédité :

```
> predict(model,type="response")
        1         2
0.8125000 0.1721311
```

Commentaire

La maladie étant rare, les probabilités d'être atteint paraissent anormalement élevées...

Cause

- L'échantillon ne parait pas être représentatif de la population.
- Les personnes atteintes sont en effet sur-représentées.
- C'est souvent le cas dans les études cas-témoin, on essaie de rééquilibrer les proportions de personnes atteintes et non atteintes dans l'échantillon.

Pourquoi rééquilibrer ?

- On considère les modèles logistiques

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et $\beta_1 = 1$ pour le modèle 1 et $\beta_1 = 5$ pour le modèle 2.

- On génère $B = 200$ échantillons $(x_1, y_1), \ldots, (x_n, y_n)$ et on s'intéresse à la distribution de l'emv $\hat{\beta}_1$ de β_1 pour les deux modèles.

```
> beta0 <- 1
> beta1 <- 5
> B <- 200
> beta <- rep(0,B)
> X <- seq(0.01,0.99,length=100)
> P <- exp(beta0+beta1*X)/(1+exp(beta0+beta1*X))
> for (i in 1:B){
+   Y <- rbinom(n,1,P)
+   model <- glm(Y~X,family=binomial)
+   beta[i] <- coef(model)[2]
+ }
```

Pourquoi rééquilibrer ?

- On considère les modèles logistiques

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et $\beta_1 = 1$ pour le modèle 1 et $\beta_1 = 5$ pour le modèle 2.

- On génère $B = 200$ échantillons $(x_1, y_1), \ldots, (x_n, y_n)$ et on s'intéresse à la distribution de l'emv $\hat{\beta}_1$ de β_1 pour les deux modèles.

```
> beta0 <- 1
> beta1 <- 5
> B <- 200
> beta <- rep(0,B)
> X <- seq(0.01,0.99,length=100)
> P <- exp(beta0+beta1*X)/(1+exp(beta0+beta1*X))
> for (i in 1:B){
+   Y <- rbinom(n,1,P)
+   model <- glm(Y~X,family=binomial)
+   beta[i] <- coef(model)[2]
+ }
```

Pourquoi rééquilibrer ?

- On considère les modèles logistiques

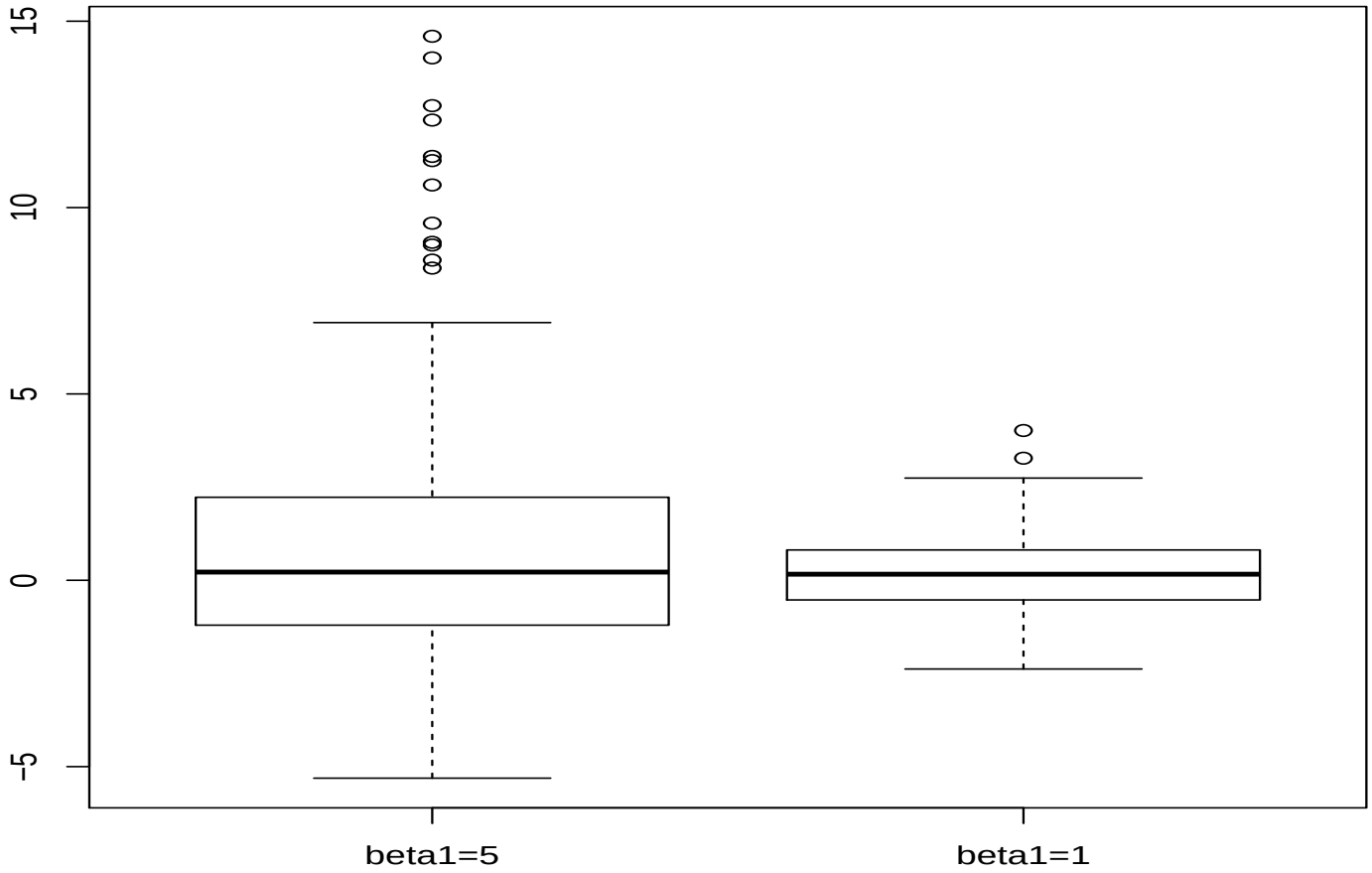
$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et $\beta_1 = 1$ pour le modèle 1 et $\beta_1 = 5$ pour le modèle 2.

- On génère $B = 200$ échantillons $(x_1, y_1), \ldots, (x_n, y_n)$ et on s'intéresse à la distribution de l'emv $\hat{\beta}_1$ de β_1 pour les deux modèles.

```
> beta0 <- 1
> beta1 <- 5
> B <- 200
> beta <- rep(0,B)
> X <- seq(0.01,0.99,length=100)
> P <- exp(beta0+beta1*X)/(1+exp(beta0+beta1*X))
> for (i in 1:B){
+   Y <- rbinom(n,1,P)
+   model <- glm(Y~X,family=binomial)
+   beta[i] <- coef(model)[2]
+ }
```

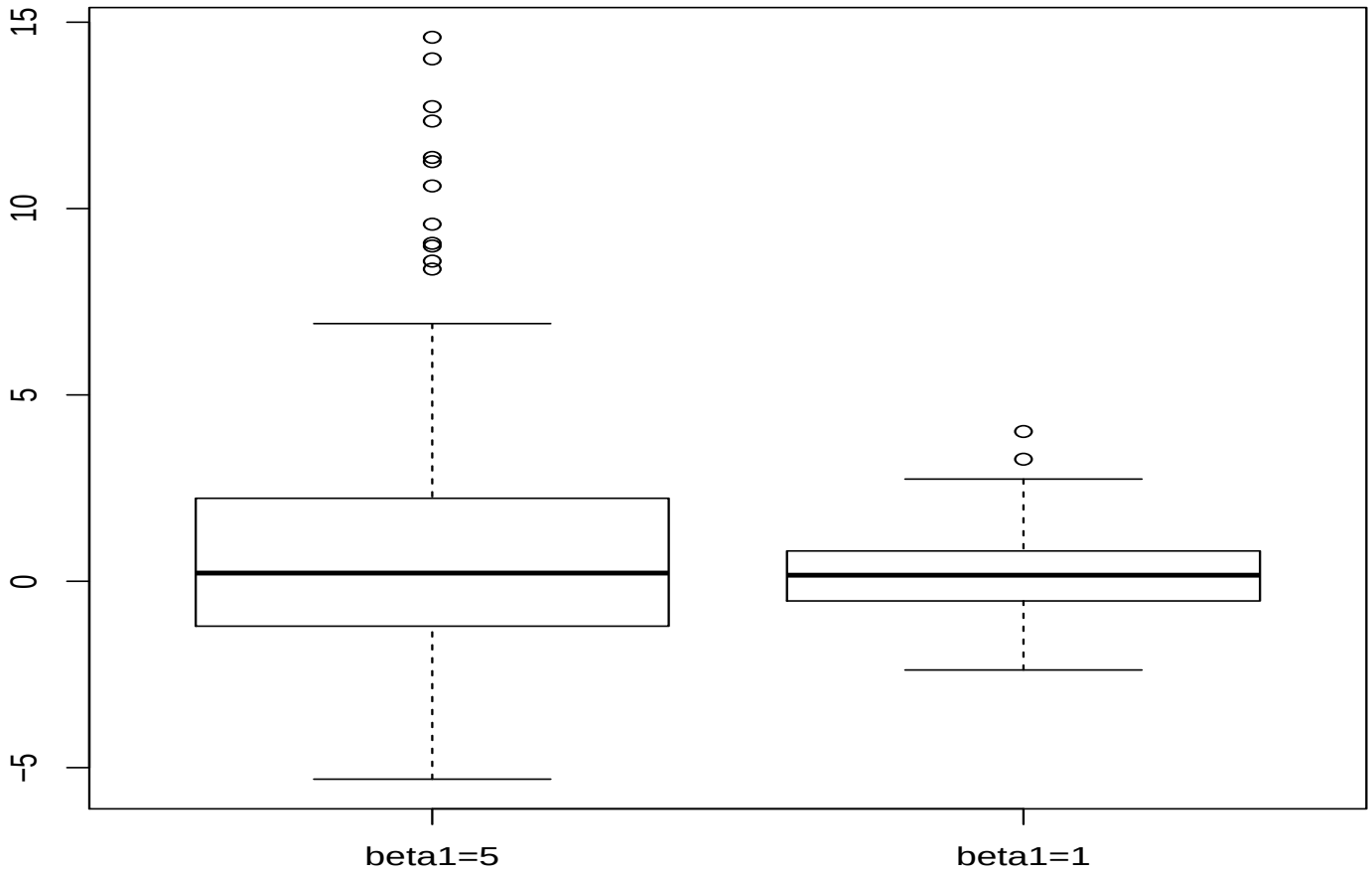
- On résume la distribution des emv de β_1 dans les deux cas à l'aide d'un boxplot.



Conclusion

Les emv ont l'air d'être sans biais, mais la variance de l'emv est plus élevée dans le cas $\beta_1 = 5$ que dans le cas $\beta_1 = 1$ (les écarts types estimés sont respectivement de 3.43 et 1.02).

- On résume la distribution des emv de β_1 dans les deux cas à l'aide d'un boxplot.

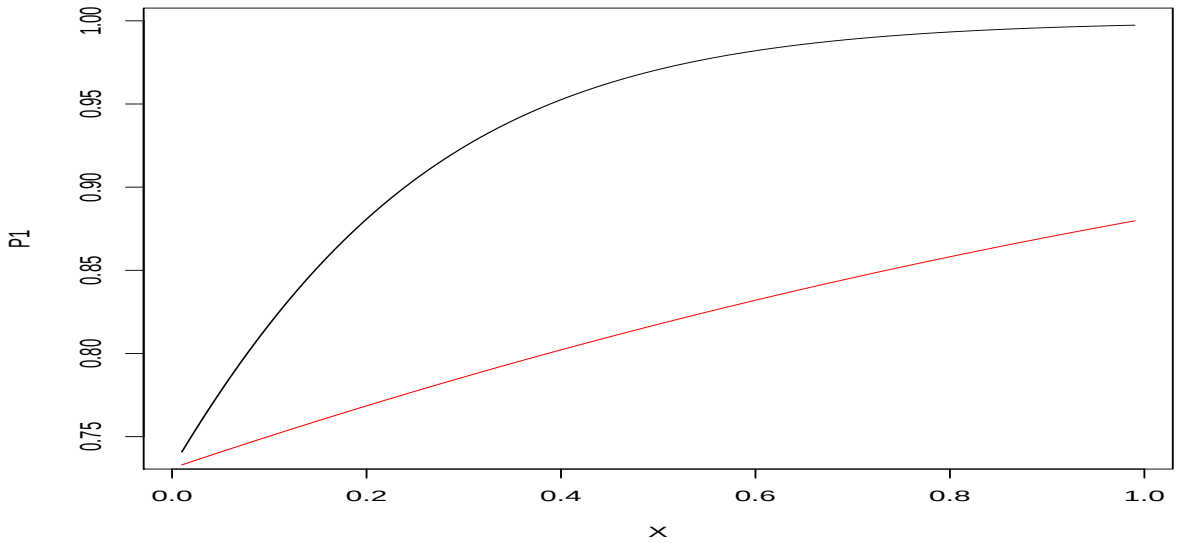


Conclusion

Les emv ont l'air d'être sans biais, mais la variance de l'emv est plus élevée dans le cas $\beta_1 = 5$ que dans le cas $\beta_1 = 1$ (les écarts types estimés sont respectivement de 3.43 et 1.02).

- Afin de comprendre cette remarque, on étudie le comportement de $p_\beta(x)$ sur $[0,1]$ pour les 2 modèles.

```
> beta11 <- 5; beta12 <- 1; P1<- exp(beta0+beta11*X)/(1+exp(beta0+beta11*X))
> P2<- exp(beta0+beta12*X)/(1+exp(beta0+beta12*X))
> plot(X,P1,type="l");lines(X,P2,col="red")
```

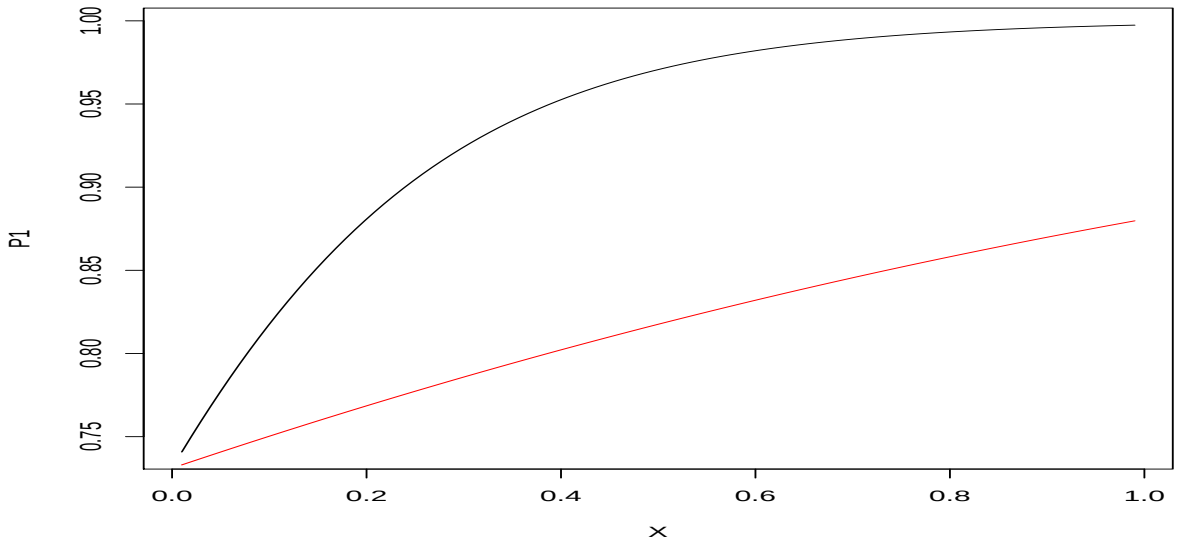


Conclusion

- Lorsque $\beta_1 = 5$, $p_\beta(x)$ se rapproche vite de 1. On aura donc une proportion de 1 dans les échantillons beaucoup plus élevée dans le cas $\beta_1 = 5$.
- En effet, sur les 200 échantillons, on a observé en moyenne : 81% de 1 dans le modèle $\beta_1 = 1$ contre 95% lorsque $\beta_1 = 5$.

- Afin de comprendre cette remarque, on étudie le comportement de $p_\beta(x)$ sur $[0,1]$ pour les 2 modèles.

```
> beta11 <- 5; beta12 <- 1; P1<- exp(beta0+beta11*X)/(1+exp(beta0+beta11*X))
> P2<- exp(beta0+beta12*X)/(1+exp(beta0+beta12*X))
> plot(X,P1,type="l");lines(X,P2,col="red")
```

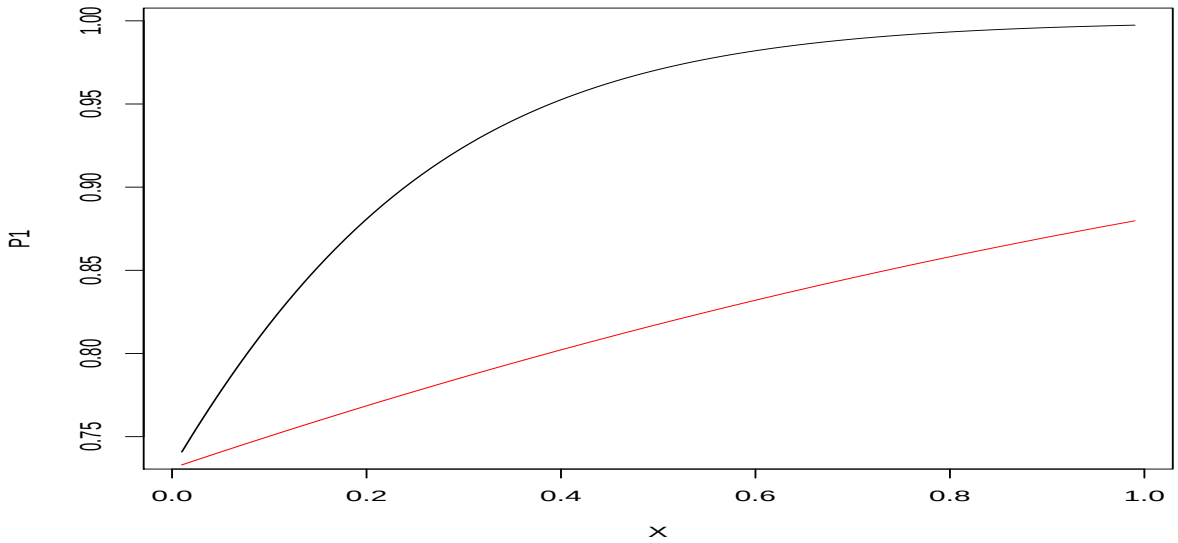


Conclusion

- Lorsque $\beta_1 = 5$, $p_\beta(x)$ se rapproche vite de 1. On aura donc une proportion de 1 dans les échantillons beaucoup plus élevée dans le cas $\beta_1 = 5$.
- En effet, sur les 200 échantillons, on a observé en moyenne : 81% de 1 dans le modèle $\beta_1 = 1$ contre 95% lorsque $\beta_1 = 5$.

- Afin de comprendre cette remarque, on étudie le comportement de $p_\beta(x)$ sur $[0,1]$ pour les 2 modèles.

```
> beta11 <- 5; beta12 <- 1; P1<- exp(beta0+beta11*X)/(1+exp(beta0+beta11*X))
> P2<- exp(beta0+beta12*X)/(1+exp(beta0+beta12*X))
> plot(X,P1,type="l");lines(X,P2,col="red")
```



Conclusion

- Lorsque $\beta_1 = 5$, $p_\beta(x)$ se rapproche vite de 1. On aura donc une proportion de 1 dans les échantillons beaucoup plus élevée dans le cas $\beta_1 = 5$.
- En effet, sur les 200 échantillons, on a observé en moyenne : 81% de 1 dans le modèle $\beta_1 = 1$ contre 95% lorsque $\beta_1 = 5$.

Justification théorique

- On rappelle que, pour n assez grand, la matrice de variance covariance de l'emv $\hat{\beta}$ du modèle logistique est

$$\mathcal{I}_n(\beta)^{-1} = (\mathbb{X}' W_\beta \mathbb{X})^{-1}$$

avec

$$W_\beta = \begin{pmatrix} p_\beta(x_1)(1-p_\beta(x_1)) & 0 & \dots & 0 \\ 0 & \ddots & & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & & p_\beta(x_n)(1-p_\beta(x_n)) \end{pmatrix}$$

Conclusion

En présence de fort déséquilibre entre les proportions de 1 et de 0, les éléments de la diagonale de W_β vont se rapprocher de 0, ce qui conduit à une augmentation de la variance des estimateurs.

Justification théorique

- On rappelle que, pour n assez grand, la matrice de variance covariance de l'emv $\hat{\beta}$ du modèle logistique est

$$\mathcal{I}_n(\beta)^{-1} = (\mathbb{X}' W_\beta \mathbb{X})^{-1}$$

avec

$$W_\beta = \begin{pmatrix} p_\beta(x_1)(1-p_\beta(x_1)) & 0 & \dots & 0 \\ 0 & \ddots & & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & & p_\beta(x_n)(1-p_\beta(x_n)) \end{pmatrix}$$

Conclusion

En présence de fort déséquilibre entre les proportions de 1 et de 0, les éléments de la diagonale de W_β vont se rapprocher de 0, ce qui conduit à une augmentation de la variance des estimateurs.

- Une solution consiste à essayer de "s'arranger" pour rééquilibrer les valeurs de Y dans l'échantillon.
- On ne peut bien entendu pas faire n'importe comment...
- Cela va forcément affecter le schéma d'échantillonnage.

Il faut le prendre compte dans l'écriture du modèle.

- Une solution consiste à essayer de "s'arranger" pour rééquilibrer les valeurs de Y dans l'échantillon.
- On ne peut bien entendu pas faire n'importe comment...
- Cela va forcément affecter le schéma d'échantillonnage.

Il faut le prendre compte dans l'écriture du modèle.

- Une solution consiste à essayer de "s'arranger" pour rééquilibrer les valeurs de Y dans l'échantillon.
- On ne peut bien entendu pas faire n'importe comment...
- Cela va forcément affecter le schéma d'échantillonnage.

Il faut le prendre compte dans l'écriture du modèle.

- Une solution consiste à essayer de "s'arranger" pour rééquilibrer les valeurs de Y dans l'échantillon.
- On ne peut bien entendu pas faire n'importe comment...
- Cela va forcément affecter le schéma d'échantillonnage.

Il faut le prendre compte dans l'écriture du modèle.

- On cherche à expliquer une variable binaire Y par une variable X à l'aide d'un modèle logistique : pour $x_i \in \mathbb{R}$, la loi de Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Problème : estimer β.
- On se place dans le cas où $\pi_1 = \mathbf{P}(Y = 1)$ est petit devant $\pi_0 = \mathbf{P}(Y = 0)$.
- On a vu que, dans ce cas, la proportion de 1 dans un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ risque d'être faible devant celle de 0, ce qui risque de nous donner des emv avec une forte variance.

Idée

On va tenter d'obtenir un échantillon avec plus de 1.

- On cherche à expliquer une variable binaire Y par une variable X à l'aide d'un modèle logistique : pour $x_i \in \mathbb{R}$, la loi de Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Problème : estimer β.
- On se place dans le cas où $\pi_1 = \mathbf{P}(Y = 1)$ est petit devant $\pi_0 = \mathbf{P}(Y = 0)$.
- On a vu que, dans ce cas, la proportion de 1 dans un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ risque d'être faible devant celle de 0, ce qui risque de nous donner des emv avec une forte variance.

Idée

On va tenter d'obtenir un échantillon avec plus de 1.

- On cherche à expliquer une variable binaire Y par une variable X à l'aide d'un modèle logistique : pour $x_i \in \mathbb{R}$, la loi de Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Problème : estimer β.
- On se place dans le cas où $\pi_1 = \mathbf{P}(Y = 1)$ est petit devant $\pi_0 = \mathbf{P}(Y = 0)$.
- On a vu que, dans ce cas, la proportion de 1 dans un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ risque d'être faible devant celle de 0, ce qui risque de nous donner des emv avec une forte variance.

Idée

On va tenter d'obtenir un échantillon avec plus de 1.

- On cherche à expliquer une variable binaire Y par une variable X à l'aide d'un modèle logistique : pour $x_i \in \mathbb{R}$, la loi de Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Problème : estimer β.
- On se place dans le cas où $\pi_1 = \mathbf{P}(Y = 1)$ est petit devant $\pi_0 = \mathbf{P}(Y = 0)$.
- On a vu que, dans ce cas, la proportion de 1 dans un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ risque d'être faible devant celle de 0, ce qui risque de nous donner des emv avec une forte variance.

Idée

On va tenter d'obtenir un échantillon avec plus de 1.

- On cherche à expliquer une variable binaire Y par une variable X à l'aide d'un modèle logistique : pour $x_i \in \mathbb{R}$, la loi de Y_i est une Bernoulli de paramètre $p_\beta(x_i)$ tel que

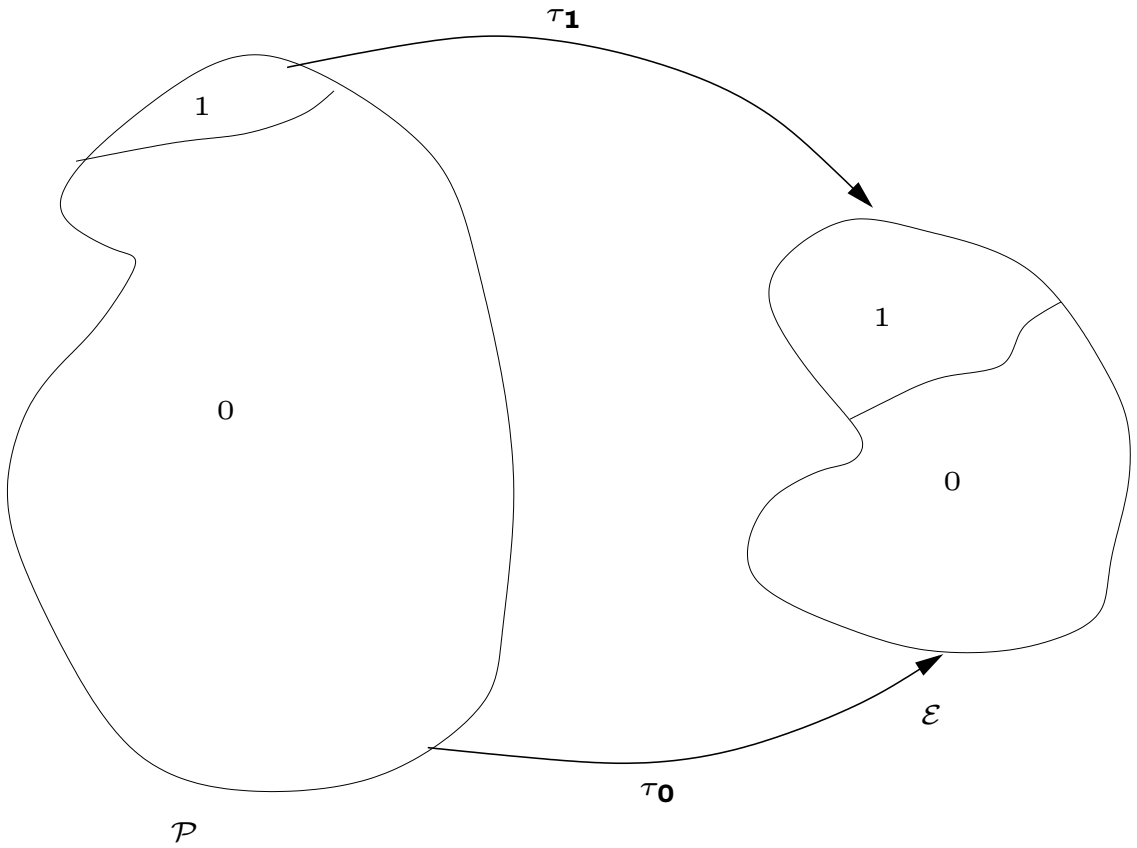
$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Problème : estimer β.
- On se place dans le cas où $\pi_1 = \mathbf{P}(Y = 1)$ est petit devant $\pi_0 = \mathbf{P}(Y = 0)$.
- On a vu que, dans ce cas, la proportion de 1 dans un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ risque d'être faible devant celle de 0, ce qui risque de nous donner des emv avec une forte variance.

Idée

On va tenter d'obtenir un échantillon avec plus de 1.

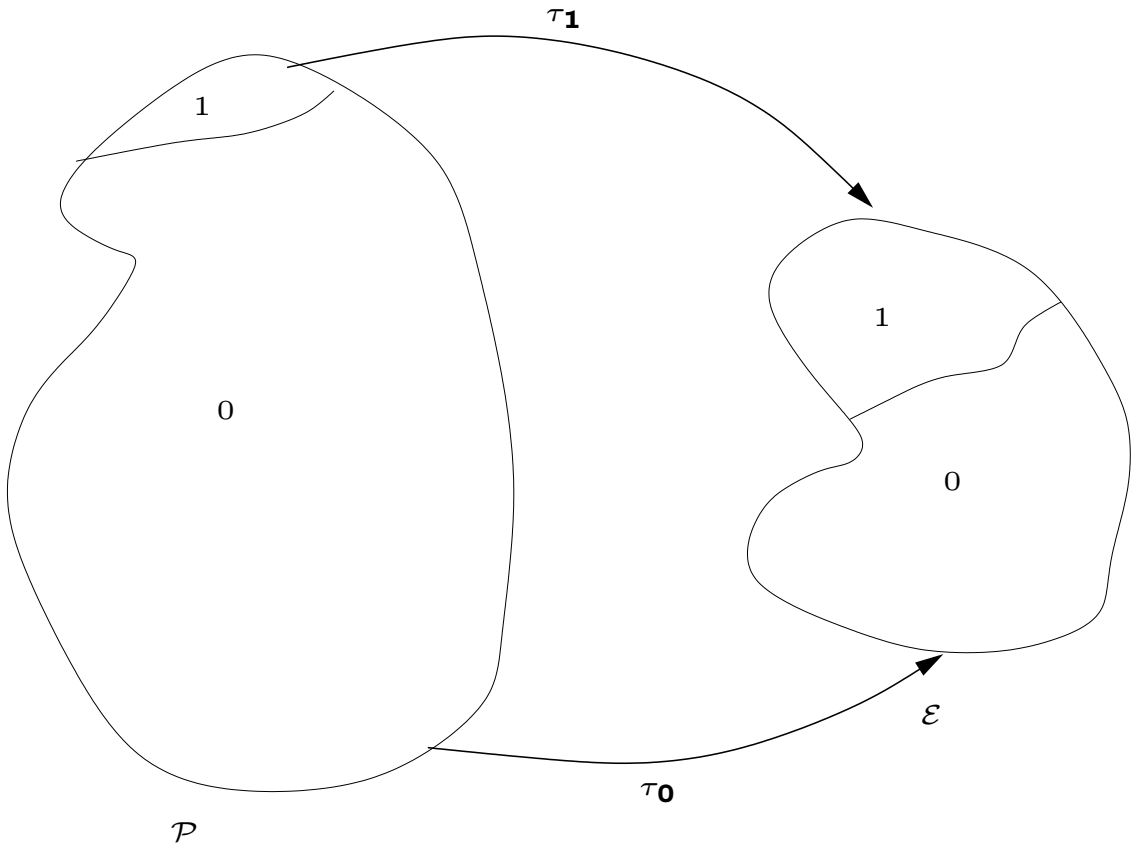
- On cherche à augmenter les 1 dans l'échantillon.



- On définit une nouvelle variable aléatoire S_i qui prend pour valeur $0, 1$ telle que

$$S_i = \begin{cases} 1 & \text{on garde l'individu } (x_i, Y_i) \text{ dans l'échantillon} \\ 0 & \text{on le supprime.} \end{cases}$$

- On cherche à augmenter les 1 dans l'échantillon.



- On définit une nouvelle variable aléatoire S_i qui prend pour valeur $0, 1$ telle que

$$S_i = \begin{cases} 1 & \text{on garde l'individu } (x_i, Y_i) \text{ dans l'échantillon} \\ 0 & \text{on le supprime.} \end{cases}$$

- On note

$$\tau_{0i} = \mathbf{P}(S_i = 1 | Y_i = 0) \quad \text{et} \quad \tau_{1i} = \mathbf{P}(S_i = 1 | Y_i = 1).$$

- Pour $i = 1, \ldots, n$, on considère le triplet (x_i, Y_i, S_i).
- Et nous allons supposer que les observations à disposition $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ sont des réalisations du triplet $(x_1, Y_1, S_1), \ldots, (x_n, Y_n, S_n)$.
- On considère le modèle logistique permettant d'expliquer Y par X à l'aide de ce triplet :

$$\operatorname{logit} p_\gamma(x_i) = \operatorname{logit} \mathbf{P}_\gamma(Y_i = 1 | S_i = 1) = \gamma_0 + \gamma_1 x_i.$$

Intérêt

Si on choisit τ_{1i} grand et τ_{0i} petit, alors les emv des paramètres γ auront a priori moins de variance que ceux des β.

- On note

$$\tau_{0i} = \mathbf{P}(S_i = 1 | Y_i = 0) \quad \text{et} \quad \tau_{1i} = \mathbf{P}(S_i = 1 | Y_i = 1).$$

- Pour $i = 1, \ldots, n$, on considère le triplet (x_i, Y_i, S_i).
- Et nous allons supposer que les observations à disposition $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ sont des réalisations du triplet $(x_1, Y_1, S_1), \ldots, (x_n, Y_n, S_n)$.
- On considère le modèle logistique permettant d'expliquer Y par X à l'aide de ce triplet :

$$\operatorname{logit} p_\gamma(x_i) = \operatorname{logit} \mathbf{P}_\gamma(Y_i = 1 | S_i = 1) = \gamma_0 + \gamma_1 x_i.$$

Intérêt

Si on choisit τ_{1i} grand et τ_{0i} petit, alors les emv des paramètres γ auront a priori moins de variance que ceux des β.

- On note

$$\tau_{0i} = \mathbf{P}(S_i = 1 | Y_i = 0) \quad \text{et} \quad \tau_{1i} = \mathbf{P}(S_i = 1 | Y_i = 1).$$

- Pour $i = 1, \ldots, n$, on considère le triplet (x_i, Y_i, S_i).
- Et nous allons supposer que les observations à disposition $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ sont des réalisations du triplet $(x_1, Y_1, S_1), \ldots, (x_n, Y_n, S_n)$.
- On considère le modèle logistique permettant d'expliquer Y par X à l'aide de ce triplet :

$$\operatorname{logit} p_\gamma(x_i) = \operatorname{logit} \mathbf{P}_\gamma(Y_i = 1 | S_i = 1) = \gamma_0 + \gamma_1 x_i.$$

Intérêt

Si on choisit τ_{1i} grand et τ_{0i} petit, alors les emv des paramètres γ auront a priori moins de variance que ceux des β.

- On note

$$\tau_{0i} = \mathbf{P}(S_i = 1 | Y_i = 0) \quad \text{et} \quad \tau_{1i} = \mathbf{P}(S_i = 1 | Y_i = 1).$$

- Pour $i = 1, \ldots, n$, on considère le triplet (x_i, Y_i, S_i).
- Et nous allons supposer que les observations à disposition $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ sont des réalisations du triplet $(x_1, Y_1, S_1), \ldots, (x_n, Y_n, S_n)$.
- On considère le modèle logistique permettant d'expliquer Y par X à l'aide de ce triplet :

$$\text{logit}\, p_\gamma(x_i) = \text{logit}\, \mathbf{P}_\gamma(Y_i = 1 | S_i = 1) = \gamma_0 + \gamma_1 x_i.$$

Intérêt

Si on choisit τ_{1i} grand et τ_{0i} petit, alors les emv des paramètres γ auront a priori moins de variance que ceux des β.

- On note

$$\tau_{0i} = \mathbf{P}(S_i = 1 | Y_i = 0) \quad \text{et} \quad \tau_{1i} = \mathbf{P}(S_i = 1 | Y_i = 1).$$

- Pour $i = 1, \ldots, n$, on considère le triplet (x_i, Y_i, S_i).
- Et nous allons supposer que les observations à disposition $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ sont des réalisations du triplet $(x_1, Y_1, S_1), \ldots, (x_n, Y_n, S_n)$.
- On considère le modèle logistique permettant d'expliquer Y par X à l'aide de ce triplet :

$$\text{logit}\, p_\gamma(x_i) = \text{logit}\, \mathbf{P}_\gamma(Y_i = 1 | S_i = 1) = \gamma_0 + \gamma_1 x_i.$$

Intérêt

Si on choisit τ_{1i} grand et τ_{0i} petit, alors les emv des paramètres γ auront a priori moins de variance que ceux des β.

Question

Quel est le lien entre β et γ ?

Théorème

On suppose que $\tau_{0i} = \tau_0$ et $\tau_{1i} = \tau_1$. Alors

$$\text{logit}\, p_\gamma(x_i) = \text{logit}\, p_\beta(x_i) + \log\left(\frac{\tau_1}{\tau_0}\right).$$

Par conséquent

$$\text{logit}\, p_\beta(x_i) = \log\left(\gamma_0 - \frac{\tau_1}{\tau_0}\right) + \gamma_1 x_i.$$

Conséquence

- Seule la constante est affectée par le biais du au schéma d'échantillonnage. On peut de plus la corriger si on connait les taux de sondage τ_0 et τ_1.
- L'emv $\hat{\gamma}_1$ est un estimateur consistant de β_1 avec a priori moins de variance que β_1.

Question

Quel est le lien entre β et γ ?

Théorème

On suppose que $\tau_{0i} = \tau_0$ et $\tau_{1i} = \tau_1$. Alors

$$\text{logit}\, p_\gamma(x_i) = \text{logit}\, p_\beta(x_i) + \log\left(\frac{\tau_1}{\tau_0}\right).$$

Par conséquent

$$\text{logit}\, p_\beta(x_i) = \log\left(\gamma_0 - \frac{\tau_1}{\tau_0}\right) + \gamma_1 x_i.$$

Conséquence

- Seule la constante est affectée par le biais du au schéma d'échantillonnage. On peut de plus la corriger si on connait les taux de sondage τ_0 et τ_1.
- L'emv $\hat{\gamma}_1$ est un estimateur consistant de β_1 avec a priori moins de variance que β_1.

Question

Quel est le lien entre β et γ ?

Théorème

On suppose que $\tau_{0i} = \tau_0$ et $\tau_{1i} = \tau_1$. Alors

$$\operatorname{logit} p_\gamma(x_i) = \operatorname{logit} p_\beta(x_i) + \log\left(\frac{\tau_1}{\tau_0}\right).$$

Par conséquent

$$\operatorname{logit} p_\beta(x_i) = \log\left(\gamma_0 - \frac{\tau_1}{\tau_0}\right) + \gamma_1 x_i.$$

Conséquence

- Seule la constante est affectée par le biais du au schéma d'échantillonnage. On peut de plus la corriger si on connait les taux de sondage τ_0 et τ_1.
- L'emv $\hat{\gamma}_1$ est un estimateur consistant de β_1 avec a priori moins de variance que β_1.

Question

Quel est le lien entre β et γ ?

Théorème

On suppose que $\tau_{0i} = \tau_0$ et $\tau_{1i} = \tau_1$. Alors

$$\text{logit}\, p_\gamma(x_i) = \text{logit}\, p_\beta(x_i) + \log\left(\frac{\tau_1}{\tau_0}\right).$$

Par conséquent

$$\text{logit}\, p_\beta(x_i) = \log\left(\gamma_0 - \frac{\tau_1}{\tau_0}\right) + \gamma_1 x_i.$$

Conséquence

- Seule la constante est affectée par le biais du au schéma d'échantillonnage. On peut de plus la corriger si on connait les taux de sondage τ_0 et τ_1.
- L'emv $\hat{\gamma}_1$ est un estimateur consistant de β_1 avec a priori moins de variance que β_1.

Exemple

- On utilise l'échantillonnage rétrospectif pour estimer les paramètres du modèle

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et et $\beta_1 = 5$ pour le modèle 2.

- On calcule les estimateurs du maximum de vraisemblance :
  1. sur un échantillon "classique" : $(x_1, y_1), \ldots, (x_n, y_n)$ tel que y_i sont générés selon des lois de Bernoulli $p_\beta(x_i)$. On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les emv calculés.
  2. sur un échantillon "rétrospectif" : $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ tels que

$$\tau_0 = \mathbf{P}(S = 1 | Y = 0) = 0.95 \quad \text{et} \quad \tau_1 = \mathbf{P}(S = 1 | Y = 1) = 0.05.$$

  On note $\hat{\gamma}_0$ et $\hat{\gamma}_1$ les emv calculés.

- On répète l'expérience 200 fois.

Exemple

- On utilise l'échantillonnage rétrospectif pour estimer les paramètres du modèle

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et et $\beta_1 = 5$ pour le modèle 2.

- On calcule les estimateurs du maximum de vraisemblance :
  1. sur un échantillon "classique" : $(x_1, y_1), \ldots, (x_n, y_n)$ tel que y_i sont générés selon des lois de Bernoulli $p_\beta(x_i)$. On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les emv calculés.
  2. sur un échantillon "rétrospectif" : $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ tels que

$$\tau_0 = \mathbf{P}(S = 1 | Y = 0) = 0.95 \quad \text{et} \quad \tau_1 = \mathbf{P}(S = 1 | Y = 1) = 0.05.$$

     On note $\hat{\gamma}_0$ et $\hat{\gamma}_1$ les emv calculés.

- On répète l'expérience 200 fois.

Exemple

- On utilise l'échantillonnage rétrospectif pour estimer les paramètres du modèle

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et et $\beta_1 = 5$ pour le modèle 2.

- On calcule les estimateurs du maximum de vraisemblance :
  1. sur un échantillon "classique" : $(x_1, y_1), \ldots, (x_n, y_n)$ tel que y_i sont générés selon des lois de Bernoulli $p_\beta(x_i)$. On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les emv calculés.
  2. sur un échantillon "rétrospectif" : $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ tels que

$$\tau_0 = \mathbf{P}(S = 1 | Y = 0) = 0.95 \quad \text{et} \quad \tau_1 = \mathbf{P}(S = 1 | Y = 1) = 0.05.$$

     On note $\hat{\gamma}_0$ et $\hat{\gamma}_1$ les emv calculés.

- On répète l'expérience 200 fois.

Exemple

- On utilise l'échantillonnage rétrospectif pour estimer les paramètres du modèle

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, 100$$

où les x_i sont uniformes sur $[0, 1]$, $\beta_0 = 1$ et et $\beta_1 = 5$ pour le modèle 2.

- On calcule les estimateurs du maximum de vraisemblance :
  1. sur un échantillon "classique" : $(x_1, y_1), \ldots, (x_n, y_n)$ tel que y_i sont générés selon des lois de Bernoulli $p_\beta(x_i)$. On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les emv calculés.
  2. sur un échantillon "rétrospectif" : $(x_1, y_1, 1), \ldots, (x_n, y_n, 1)$ tels que

$$\tau_0 = \mathbf{P}(S = 1 | Y = 0) = 0.95 \quad \text{et} \quad \tau_1 = \mathbf{P}(S = 1 | Y = 1) = 0.05.$$

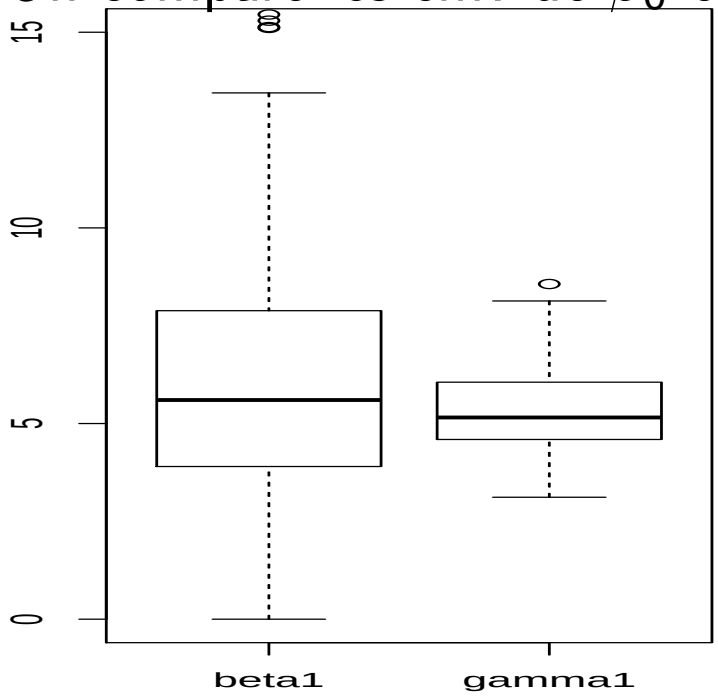
     On note $\hat{\gamma}_0$ et $\hat{\gamma}_1$ les emv calculés.

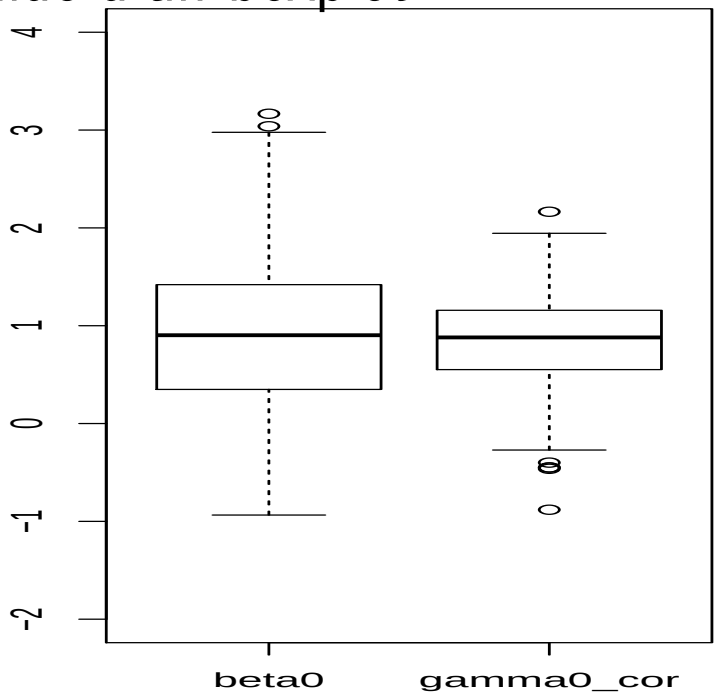
- On répète l'expérience 200 fois.

Code R

```
> n <- 100; beta0 <- 1; beta1 <- 5; B <- 200
> beta <- matrix(0,ncol=2,nrow=B); gamma <- beta
> X <- seq(0.02,0.98,length=100)
> Y <- rep(0,n)
> P <- exp(beta0+beta1*X)/(1+exp(beta0+beta1*X))
> tau <- c(0.95,0.05)
> for (i in 1:B){
+   for (j in (1:n)){
+     s <- 0
+     k <- 0
+     while (s==0){
+       y <- rbinom(1,1,P[j])
+       s <- rbinom(1,1,tau[y+1])
+     }
+     Y[j] <- y
+   }
+   Y0 <- rbinom(n,1,P)
+   model <- glm(Y~X,family=binomial)
+   model0 <- glm(Y0~X,family=binomial)
+   gamma[i,] <- coef(model)
+   beta[i,] <- coef(model0)
+ }
```

- On compare les emv de β_0 et β_1 à l'aide d'un boxplot.

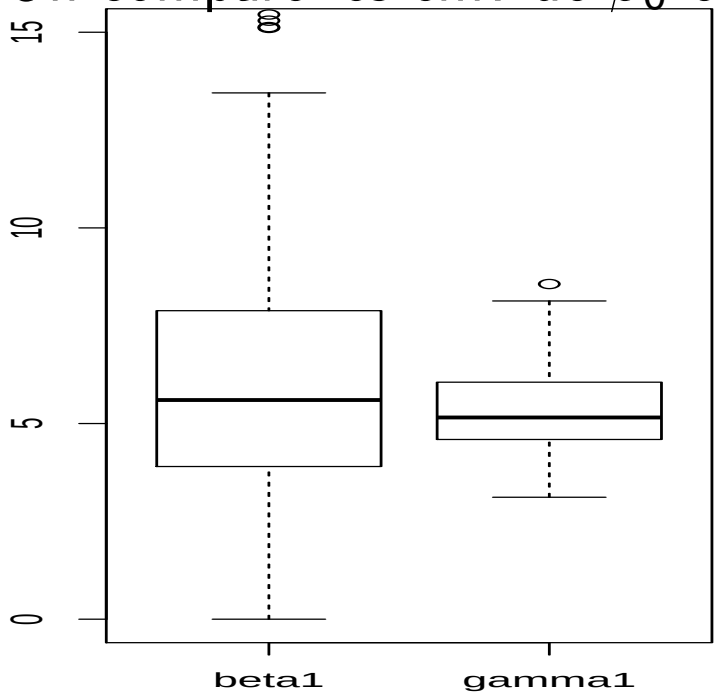


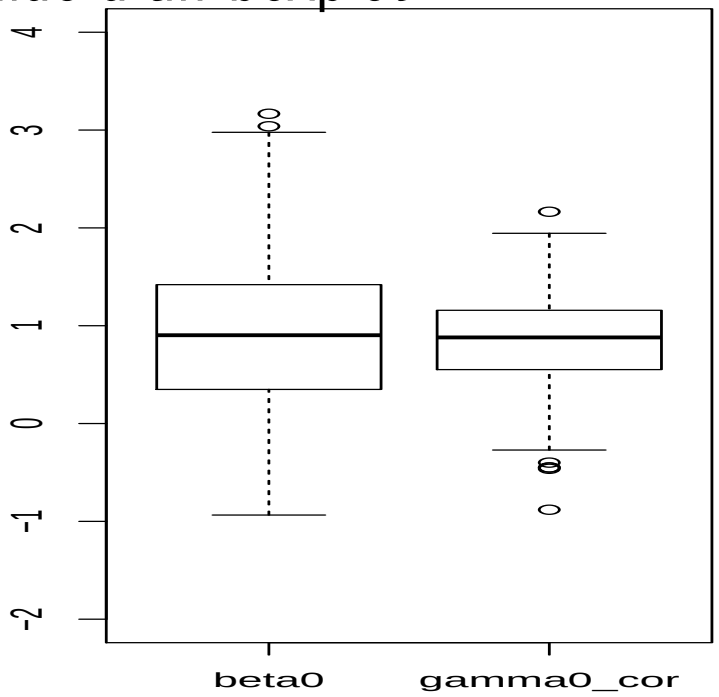


En terme de biais, les performances des estimateurs sont comparables. La variance des $\hat{\gamma}$ est par contre nettement plus petite que celle des $\hat{\beta}$.

```
> apply(beta,2,mean)
[1] 1.049845 6.290707
> apply(gamma,2,mean)
[1] -2.078106  5.314054
> apply(beta,2,sd)
[1] 1.992387 3.777083
> apply(gamma,2,sd)
[1] 0.5158525 1.0585252
```

- On compare les emv de β_0 et β_1 à l'aide d'un boxplot.





En terme de biais, les performances des estimateurs sont comparables. La variance des $\hat{\gamma}$ est par contre nettement plus petite que celle des $\hat{\beta}$.

```
> apply(beta,2,mean)
[1] 1.049845 6.290707
> apply(gamma,2,mean)
[1] -2.078106  5.314054
> apply(beta,2,sd)
[1] 1.992387 3.777083
> apply(gamma,2,sd)
[1] 0.5158525 1.0585252
```

En pratique

Cette propriété remarquable du modèle logistique dans le cadre d'un échantillonnage rétrospectif peut être appliquée dans (au moins) deux cas.

1. Les études cas-témoins (très utilisées en épidémiologie). On souhaite par exemple mesurer l'importance d'un caractère sur une pathologie. On construit alors l'échantillon en sélectionnant
   - un nombre n_1 fixé de patients atteints (cas) ;
   - un nombre n_0 fixé de patients sains (témoin).
2. Lorsque que l'on dispose d'une grande base de données dans lesquels les individus 1 sont sous représentés. On construit alors une deuxième base de données (plus petite) en donnant un poids plus élevé aux individus 1 pour être dans la seconde base ($\tau_1 > \tau_0$).

Remarque

Dans le second cas, rien ne garantit que les estimateurs calculés sur la petite base soient plus performants que ceux calculés sur la base initiale.

En pratique

Cette propriété remarquable du modèle logistique dans le cadre d'un échantillonnage rétrospectif peut être appliquée dans (au moins) deux cas.

1. Les études cas-témoins (très utilisées en épidémiologie). On souhaite par exemple mesurer l'importance d'un caractère sur une pathologie. On construit alors l'échantillon en sélectionnant
   - un nombre n_1 fixé de patients atteints (cas) ;
   - un nombre n_0 fixé de patients sains (témoin).
2. Lorsque que l'on dispose d'une grande base de données dans lesquels les individus 1 sont sous représentés. On construit alors une deuxième base de données (plus petite) en donnant un poids plus élevé aux individus 1 pour être dans la seconde base ($\tau_1 > \tau_0$).

Remarque

Dans le second cas, rien ne garantit que les estimateurs calculés sur la petite base soient plus performants que ceux calculés sur la base initiale.

En pratique

Cette propriété remarquable du modèle logistique dans le cadre d'un échantillonnage rétrospectif peut être appliquée dans (au moins) deux cas.

1. Les études cas-témoins (très utilisées en épidémiologie). On souhaite par exemple mesurer l'importance d'un caractère sur une pathologie. On construit alors l'échantillon en sélectionnant
   - un nombre n_1 fixé de patients atteints (cas) ;
   - un nombre n_0 fixé de patients sains (témoin).
2. Lorsque que l'on dispose d'une grande base de données dans lesquels les individus 1 sont sous représentés. On construit alors une deuxième base de données (plus petite) en donnant un poids plus élevé aux individus 1 pour être dans la seconde base ($\tau_1 > \tau_0$).

Remarque

Dans le second cas, rien ne garantit que les estimateurs calculés sur la petite base soient plus performants que ceux calculés sur la base initiale.

Sixième partie VI

Grande dimension : régression logistique pénalisée

- Lorsque le nombre de variables d est grand, les estimateurs du maximum de vraisemblance du modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_d x_d$$

possèdent généralement une grande variance.

Idée des méthodes pénalisés

- Contraindre la valeur des estimateurs du maximum de vraisemblance de manière à réduire la variance (quitte à augmenter un peu le biais).
- Comment ? En imposant une contrainte sur la valeur des estimateurs du MV :

$$\hat{\beta}^{pen} = \underset{\beta}{\text{argmax}}\, \mathcal{L}_n(\beta) = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \{y_i x_i' \beta - \log(1 + \exp(x_i' \beta))\}$$

$$\text{sous la contrainte } \|\beta\|_? \leq t.$$

- Lorsque le nombre de variables d est grand, les estimateurs du maximum de vraisemblance du modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_d x_d$$

possèdent généralement une grande variance.

Idée des méthodes pénalisés

- Contraindre la valeur des estimateurs du maximum de vraisemblance de manière à réduire la variance (quitte à augmenter un peu le biais).
- Comment ? En imposant une contrainte sur la valeur des estimateurs du MV :

$$\hat{\beta}^{pen} = \underset{\beta}{\text{argmax}}\, \mathcal{L}_n(\beta) = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \{y_i x_i' \beta - \log(1 + \exp(x_i' \beta))\}$$

$$\text{sous la contrainte } \|\beta\|_? \leq t.$$

- Lorsque le nombre de variables d est grand, les estimateurs du maximum de vraisemblance du modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_d x_d$$

possèdent généralement une grande variance.

Idée des méthodes pénalisés

- Contraindre la valeur des estimateurs du maximum de vraisemblance de manière à réduire la variance (quitte à augmenter un peu le biais).
- Comment ? En imposant une contrainte sur la valeur des estimateurs du MV :

$$\hat{\beta}^{pen} = \underset{\beta}{\text{argmax}}\, \mathcal{L}_n(\beta) = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\}$$

$$\text{sous la contrainte } \|\beta\|_? \leq t.$$

Questions

- Quelle norme utiliser pour la contrainte ?
- Existence/unicité des estimateurs ? Solutions explicites du problème d'optimisation ?
- Comment choisir t ?
  - t grand $\Longrightarrow$ estimateurs contraints (proche de 0) ;
  - t petit $\Longrightarrow$ estimateurs du maximum de vraisemblance (non pénalisés).

Questions

- Quelle norme utiliser pour la contrainte ?
- Existence/unicité des estimateurs ? Solutions explicites du problème d'optimisation ?
- Comment choisir t ?
  - t grand $\Longrightarrow$ estimateurs contraints (proche de 0) ;
  - t petit $\Longrightarrow$ estimateurs du maximum de vraisemblance (non pénalisés).

Questions

- Quelle norme utiliser pour la contrainte ?
- Existence/unicité des estimateurs ? Solutions explicites du problème d'optimisation ?
- Comment choisir t ?
  - t grand $\Longrightarrow$ estimateurs contraints (proche de 0) ;
  - t petit $\Longrightarrow$ estimateurs du maximum de vraisemblance (non pénalisés).

- La régression ridge consiste à maximiser la vraisemblance pénalisée par la norme 2 des coefficients.

Définition

1. *Les estimateurs ridge* $\hat{\beta}^R$ *s'obtiennent en maximisant*

$$\sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} \quad sous\ la\ contrainte \quad \sum_{j=1}^{d} \beta_j^2 \leq t \tag{7}$$

2. *ou de façon équivalente*

$$\hat{\beta}^R = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} - \lambda \sum_{j=1}^{d} \beta_j^2 \right\}. \tag{8}$$

Quelques remarques

- Les définitions (9) et (10) sont équivalentes dans le sens où pour tout t il existe un unique μ tels que les solutions aux deux problèmes d'optimisation coïncident.
- La constante β_0 n'entre généralement pas dans la pénalité.
- L'estimateur dépend bien entendu du paramètre t (ou λ) : $\hat{\beta}^R = \hat{\beta}^R(t) = \hat{\beta}^R(\lambda)$.
- Le plus souvent, les variables explicatives sont réduites pour éviter les problèmes d'échelle dans la pénalité.

Quelques remarques

- Les définitions (9) et (10) sont équivalentes dans le sens où pour tout t il existe un unique μ tels que les solutions aux deux problèmes d'optimisation coïncident.
- La constante β_0 n'entre généralement pas dans la pénalité.
- L'estimateur dépend bien entendu du paramètre t (ou λ) : $\hat{\beta}^R = \hat{\beta}^R(t) = \hat{\beta}^R(\lambda)$.
- Le plus souvent, les variables explicatives sont réduites pour éviter les problèmes d'échelle dans la pénalité.

Quelques remarques

- Les définitions (9) et (10) sont équivalentes dans le sens où pour tout t il existe un unique μ tels que les solutions aux deux problèmes d'optimisation coïncident.
- La constante β_0 n'entre généralement pas dans la pénalité.
- L'estimateur dépend bien entendu du paramètre t (ou λ) : $\hat{\beta}^R = \hat{\beta}^R(t) = \hat{\beta}^R(\lambda)$.
- Le plus souvent, les variables explicatives sont réduites pour éviter les problèmes d'échelle dans la pénalité.

Quelques remarques

- Les définitions (9) et (10) sont équivalentes dans le sens où pour tout t il existe un unique μ tels que les solutions aux deux problèmes d'optimisation coïncident.
- La constante β_0 n'entre généralement pas dans la pénalité.
- L'estimateur dépend bien entendu du paramètre t (ou λ) : $\hat{\beta}^R = \hat{\beta}^R(t) = \hat{\beta}^R(\lambda)$.
- Le plus souvent, les variables explicatives sont réduites pour éviter les problèmes d'échelle dans la pénalité.

Un exemple

- On reprend les données sur la maladie cardiovasculaire.

```
> data(SAheart,package="bestglm")
> SAheart[1:5,]
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
```

- Il existe plusieurs fonctions et packages qui permettent de faire de la régression pénalisée sur R. Nous présentons ici `glmnet`.

Un exemple

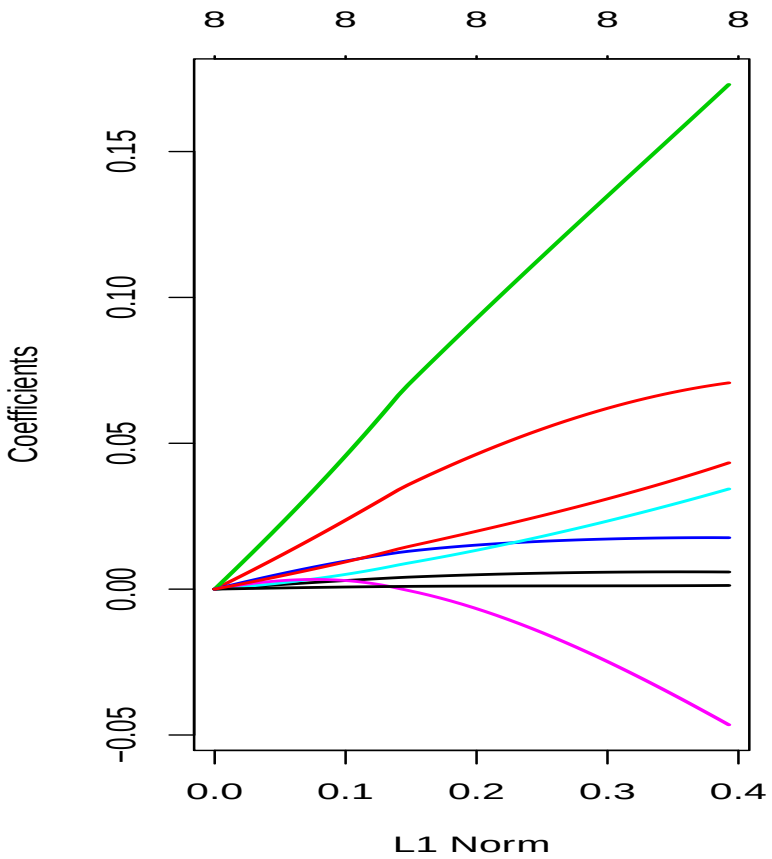
- On reprend les données sur la maladie cardiovasculaire.

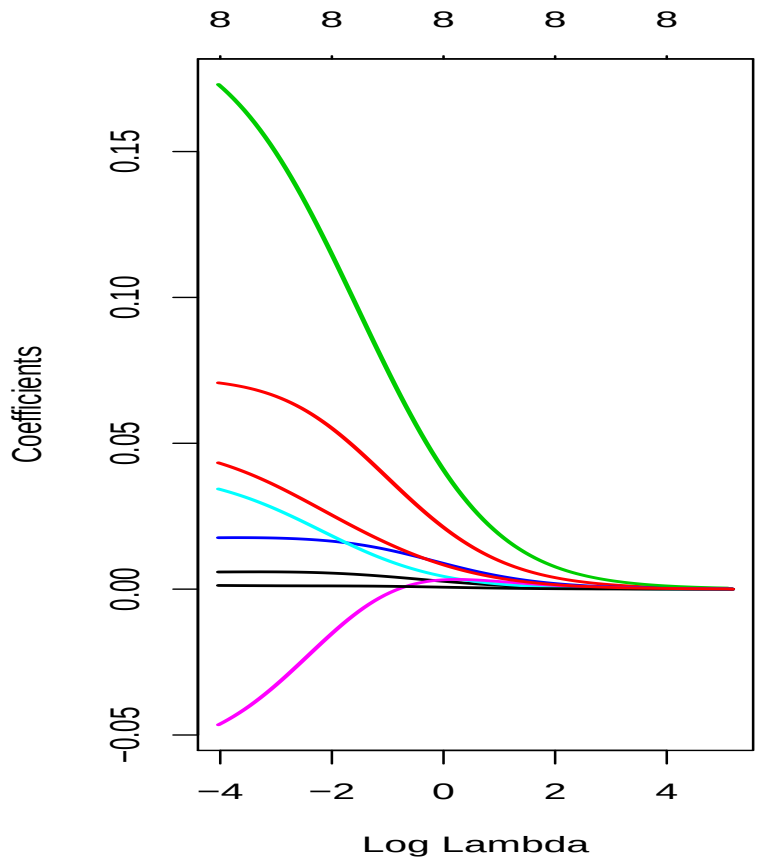
```
> data(SAheart,package="bestglm")
> SAheart[1:5,]
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
```

- Il existe plusieurs fonctions et packages qui permettent de faire de la régression pénalisée sur R. Nous présentons ici `glmnet`.

Le coin R

```
> SAheart1 <- data.matrix(SAheart)
> SAheart1 <- SAheart1[,-5]
> reg.ridge <- glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=0)
> plot(reg.ridge,label=TRUE,lwd=2)
> plot(reg.ridge,xvar="lambda",label=TRUE,lwd=2)
```





Choix de λ

- Il est crucial : si $\lambda \approx 0$ alors $\hat{\beta}^R \approx \hat{\beta}^{MV}$, si λ "grand" alors $\hat{\beta}^R \approx 0$.
- Le choix de λ se fait le plus souvent de façon "classique" :
  1. Estimation d'un critère de choix de modèle pour toutes les valeurs de λ ;
  2. Choix du λ qui minimise le critère estimé.
- Exemple : la fonction cv.glmnet choisit la valeur de λ qui minimise la déviance estimée par validation croisée.

Choix de λ

- Il est crucial : si $\lambda \approx 0$ alors $\hat{\beta}^R \approx \hat{\beta}^{MV}$, si λ "grand" alors $\hat{\beta}^R \approx 0$.
- Le choix de λ se fait le plus souvent de façon "classique" :
  1. Estimation d'un critère de choix de modèle pour toutes les valeurs de λ ;
  2. Choix du λ qui minimise le critère estimé.
- Exemple : la fonction cv.glmnet choisit la valeur de λ qui minimise la déviance estimée par validation croisée.

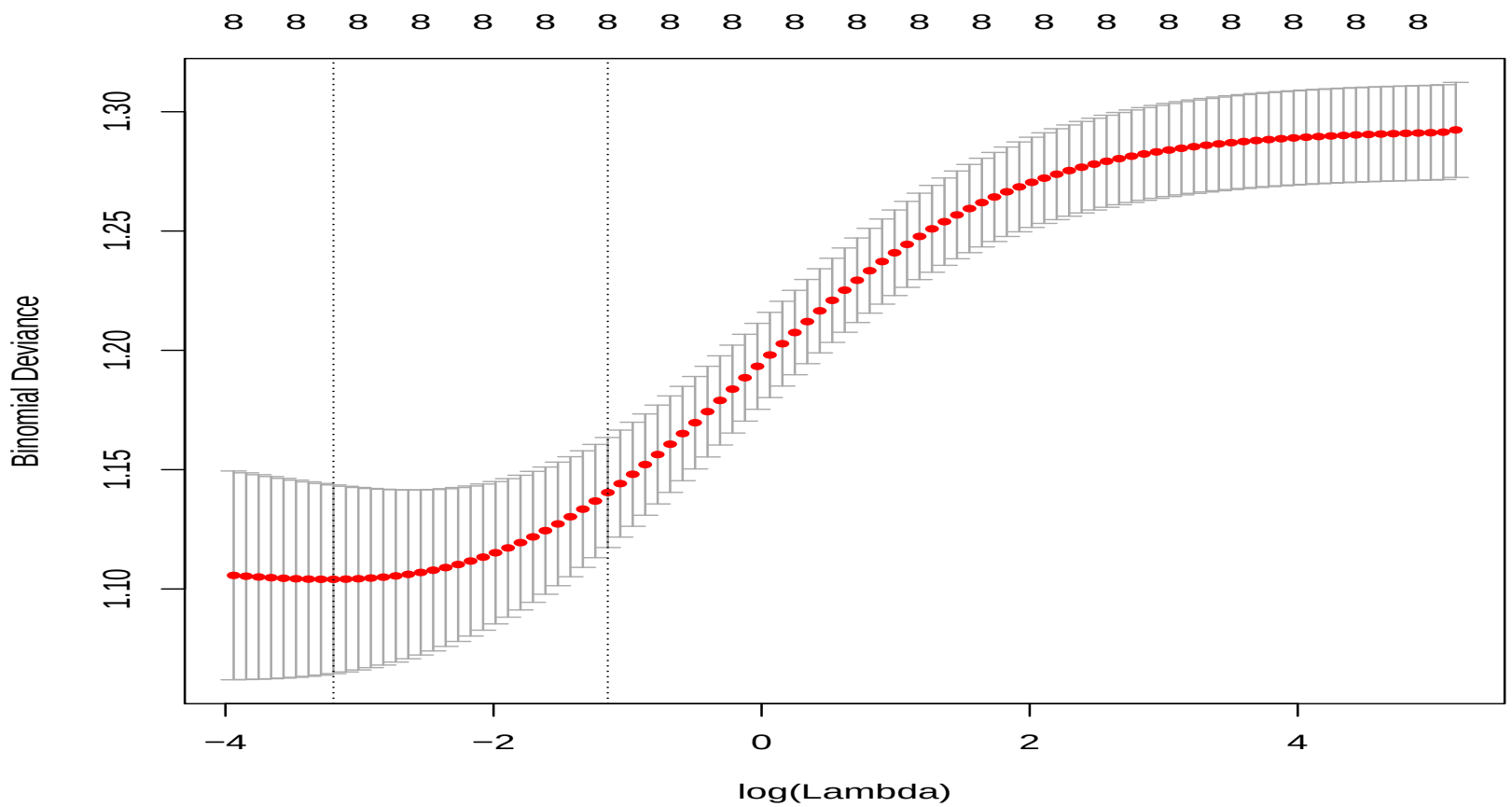
Choix de λ

- Il est crucial : si $\lambda \approx 0$ alors $\hat{\beta}^R \approx \hat{\beta}^{MV}$, si λ "grand" alors $\hat{\beta}^R \approx 0$.
- Le choix de λ se fait le plus souvent de façon "classique" :
  1. Estimation d'un critère de choix de modèle pour toutes les valeurs de λ ;
  2. Choix du λ qui minimise le critère estimé.
- Exemple : la fonction cv.glmnet choisit la valeur de λ qui minimise la déviance estimée par validation croisée.

Choix de λ

- Il est crucial : si $\lambda \approx 0$ alors $\hat{\beta}^R \approx \hat{\beta}^{MV}$, si λ "grand" alors $\hat{\beta}^R \approx 0$.
- Le choix de λ se fait le plus souvent de façon "classique" :
  1. Estimation d'un critère de choix de modèle pour toutes les valeurs de λ ;
  2. Choix du λ qui minimise le critère estimé.
- Exemple : la fonction cv.glmnet choisit la valeur de λ qui minimise la déviance estimée par validation croisée.

```
> reg.cvridge <- cv.glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=0)
> bestlam <- reg.cvridge$lambda.min
> bestlam
[1] 0.04099545
> plot(reg.cvridge)
```

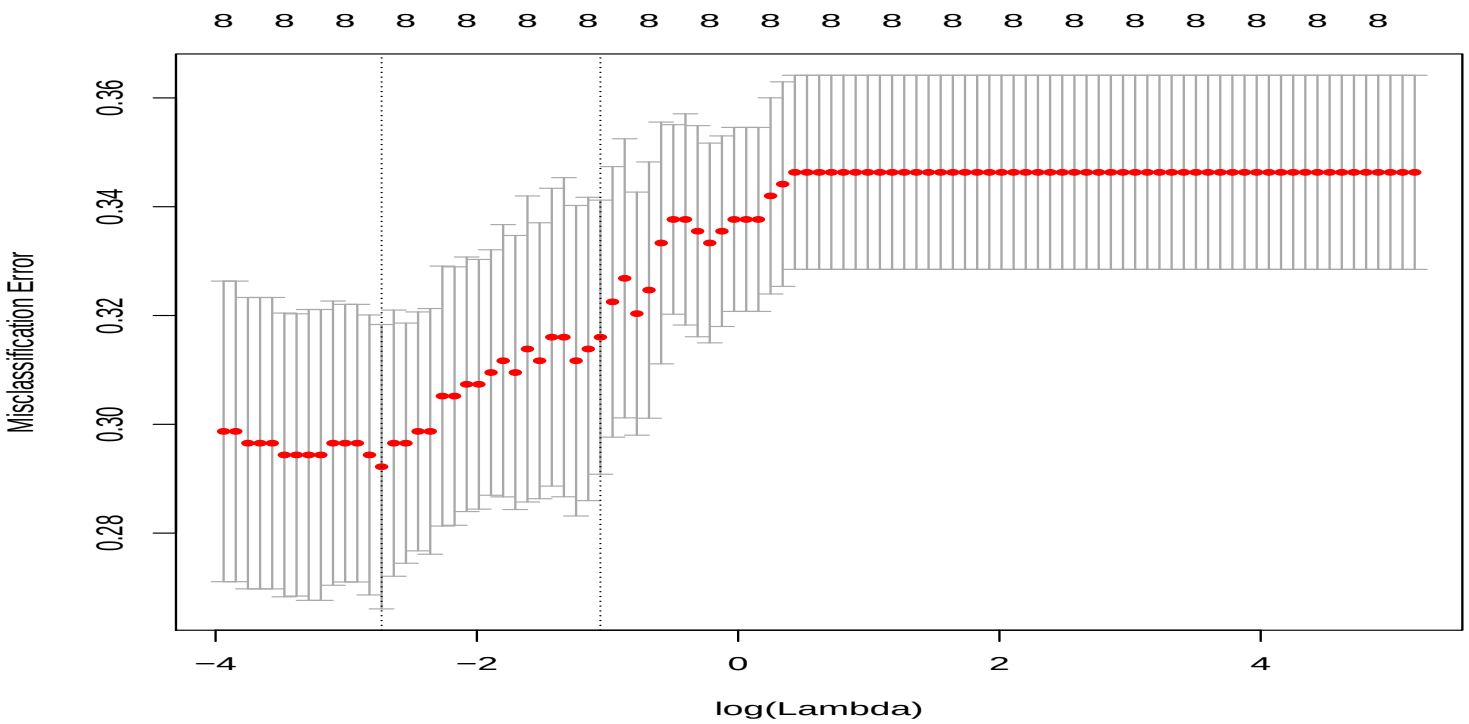


- Pour changer de critère, il suffit de modifier l'argument `type.measure` dans cv.glmnet.
- Par exemple, pour le taux de mauvais classement

```
> reg.cvridge <- cv.glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=0,
      type.measure="class")
> bestlam <- reg.cvridge$lambda.min
> bestlam
[1] 0.06527635
> plot(reg.cvridge)
```

- Pour changer de critère, il suffit de modifier l'argument `type.measure` dans cv.glmnet.
- Par exemple, pour le taux de mauvais classement

```
> reg.cvridge <- cv.glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=0,
      type.measure="class")
> bestlam <- reg.cvridge$lambda.min
> bestlam
[1] 0.06527635
> plot(reg.cvridge)
```



- La régression lasso consiste à maximiser la vraisemblance pénalisée par la norme 1 des coefficients.

Définition ([Tibshirani, 1996])

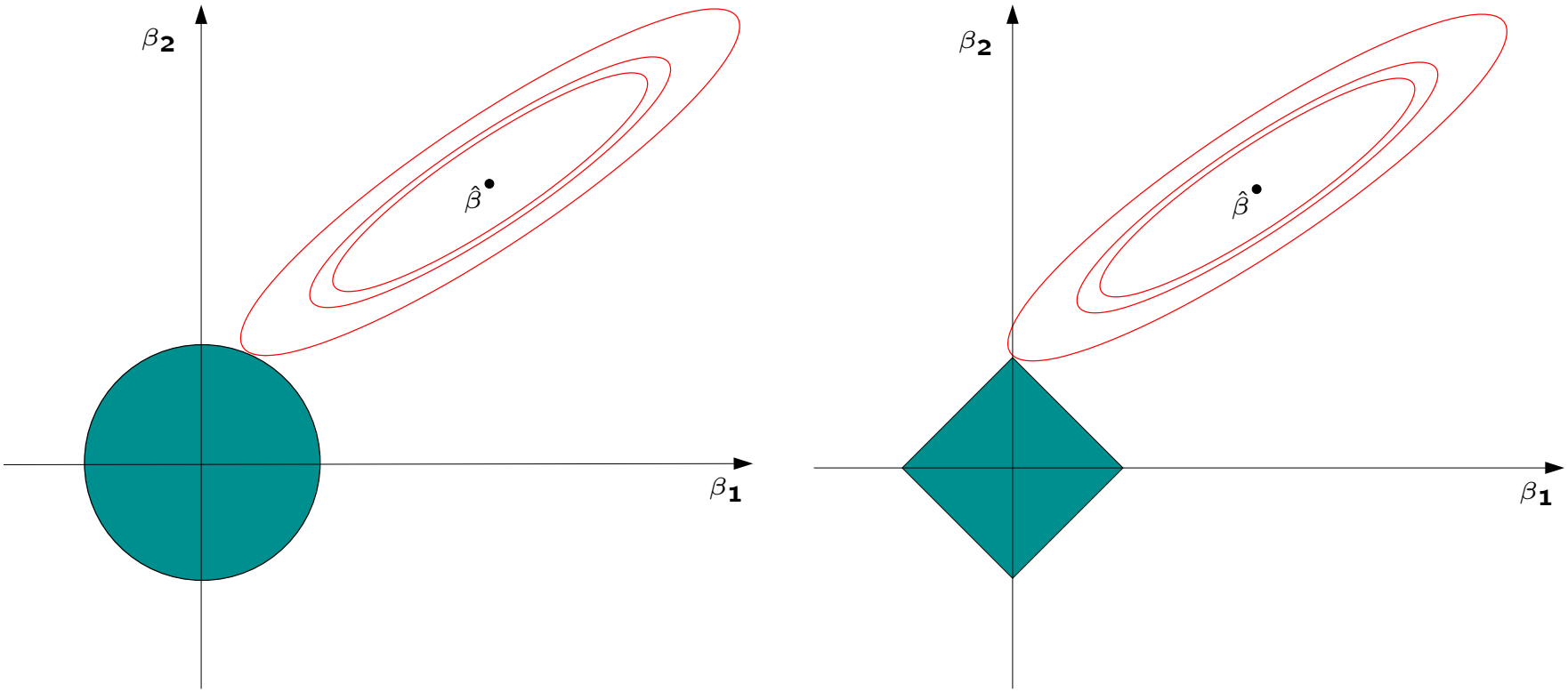
1. *Les estimateurs lasso $\hat{\beta}^L$ s'obtiennent en maximisant*

$$\sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} \quad sous\ la\ contrainte \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{9}$$

2. *ou de façon équivalente*

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} - \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{10}$$

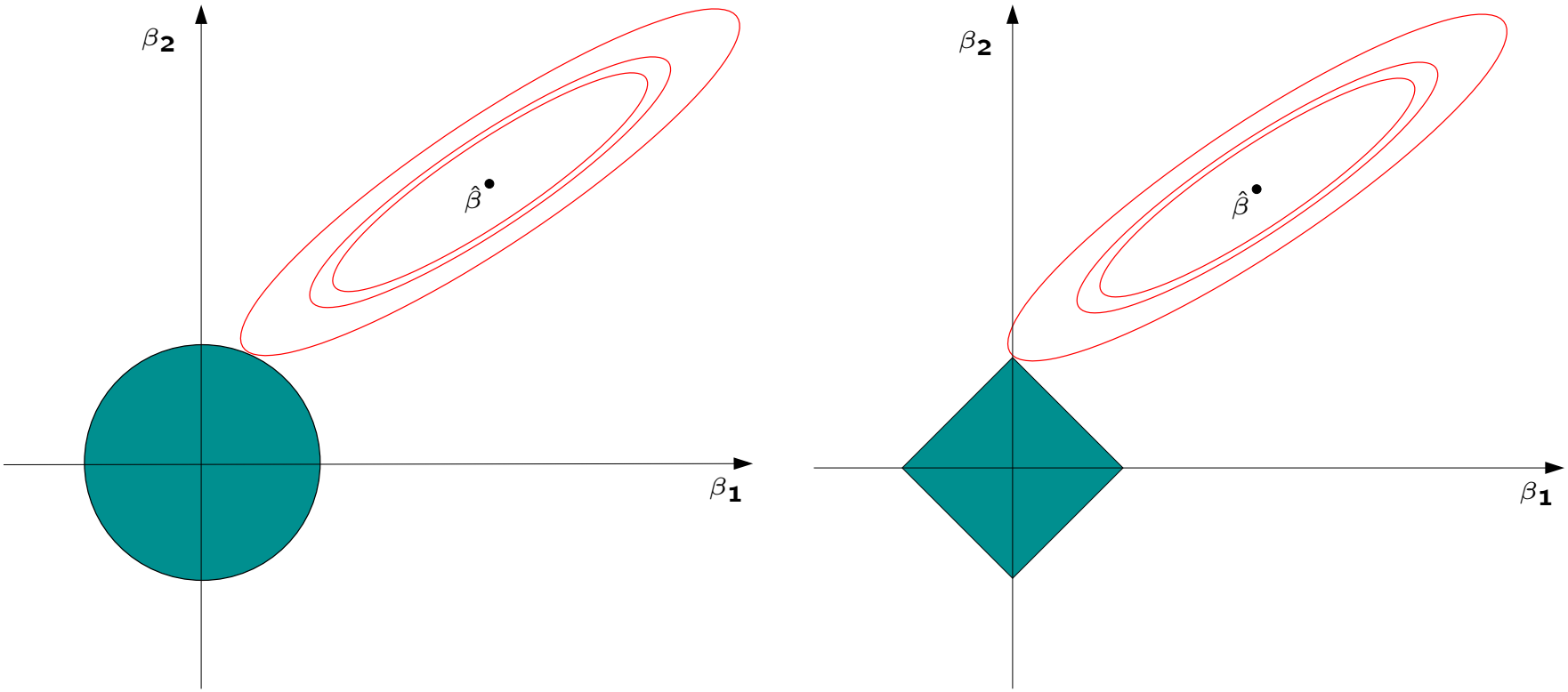
Comparaison ridge/lasso



Ces approches reviennent (d'une certaine façon) à projeter l'estimateur du MV sur les boules unités associées à

1. la norme 2 pour la régression ridge ;
2. la norme 1 pour le lasso.

Comparaison ridge/lasso



Ces approches reviennent (d'une certaine façon) à projeter l'estimateur du MV sur les boules unités associées à

1. la norme 2 pour la régression ridge ;
2. la norme 1 pour le lasso.

Quelques remarques

- Comme pour la régression ridge :
  - on préfère souvent réduire la matrice de design avant d'effectuer la régression lasso ;
  - Le choix de λ est crucial (il est le plus souvent sélectionné en minimisant un critère empirique).
  - $\lambda \nearrow \Longrightarrow$ biais $\nearrow$ et variance $\searrow$ et réciproquement lorsque $\lambda \searrow$.
- MAIS, contrairement à ridge : $\lambda \nearrow \Longrightarrow$ le nombre de coefficients nuls augmente ([Bühlmann and van de Geer, 2011]).

Quelques remarques

- Comme pour la régression ridge :
  - on préfère souvent réduire la matrice de design avant d'effectuer la régression lasso ;
  - Le choix de λ est crucial (il est le plus souvent sélectionné en minimisant un critère empirique).
  - $\lambda \nearrow \Longrightarrow$ biais $\nearrow$ et variance $\searrow$ et réciproquement lorsque $\lambda \searrow$.
- MAIS, contrairement à ridge : $\lambda \nearrow \Longrightarrow$ le nombre de coefficients nuls augmente ([Bühlmann and van de Geer, 2011]).

Quelques remarques

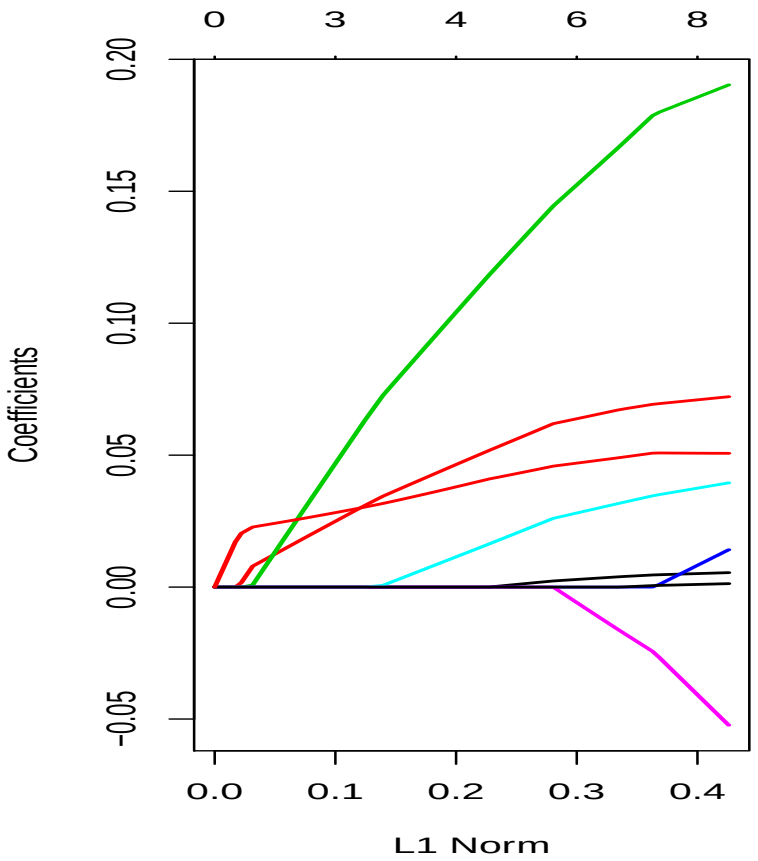
- Comme pour la régression ridge :
  - on préfère souvent réduire la matrice de design avant d'effectuer la régression lasso ;
  - Le choix de λ est crucial (il est le plus souvent sélectionné en minimisant un critère empirique).
  - $\lambda \nearrow \Longrightarrow$ biais $\nearrow$ et variance $\searrow$ et réciproquement lorsque $\lambda \searrow$.
- MAIS, contrairement à ridge : $\lambda \nearrow \Longrightarrow$ le nombre de coefficients nuls augmente ([Bühlmann and van de Geer, 2011]).

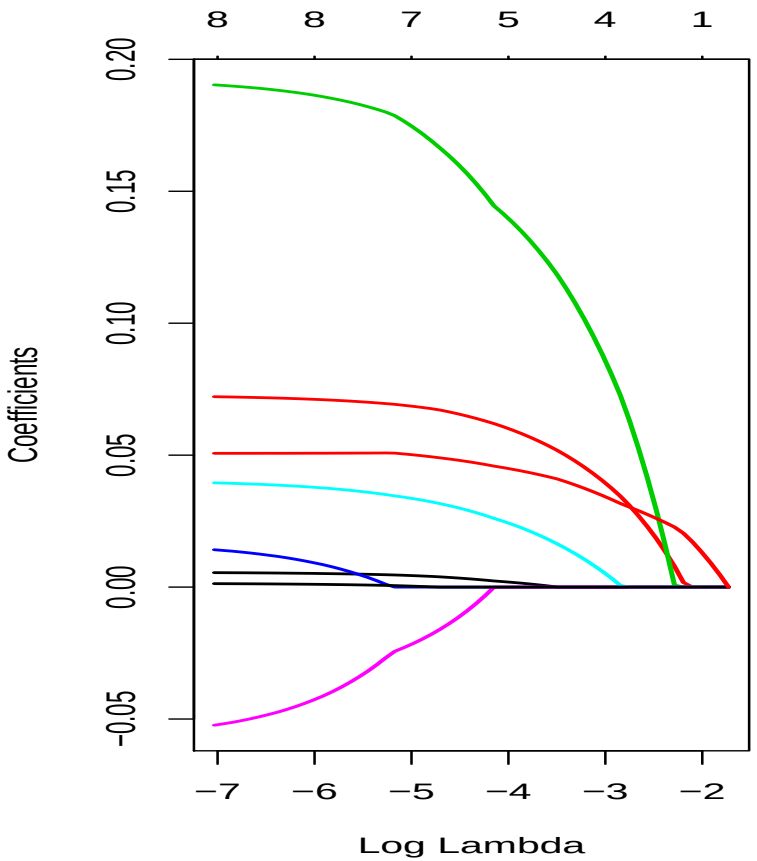
Quelques remarques

- Comme pour la régression ridge :
  - on préfère souvent réduire la matrice de design avant d'effectuer la régression lasso ;
  - Le choix de λ est crucial (il est le plus souvent sélectionné en minimisant un critère empirique).
  - $\lambda \nearrow \Longrightarrow$ biais $\nearrow$ et variance $\searrow$ et réciproquement lorsque $\lambda \searrow$.
- MAIS, contrairement à ridge : $\lambda \nearrow \Longrightarrow$ le nombre de coefficients nuls augmente ([Bühlmann and van de Geer, 2011]).

Le coin R

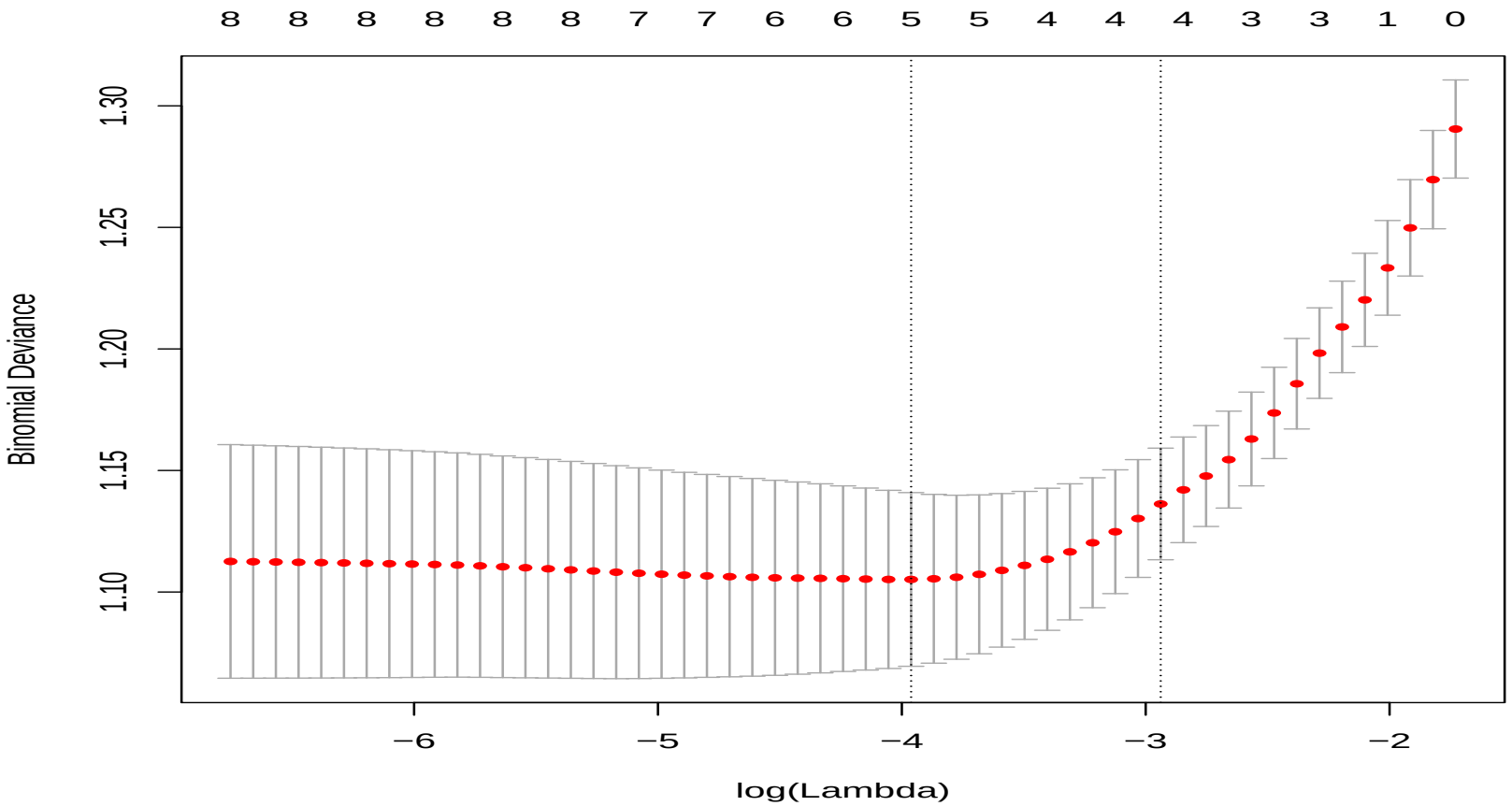
```
> reg.lasso <- glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=1)
> plot(reg.lasso,label=TRUE,lwd=2)
> plot(reg.lasso,xvar="lambda",label=TRUE,lwd=2)
```





Sélection de λ

```
> reg.cvlasso <- cv.glmnet(SAheart1[,1:8],SAheart1[,9],family="binomial",alpha=1)
> bestlam <- reg.cvlasso$lambda.min
> bestlam
[1] 0.0190284
> plot(reg.cvlasso)
```



Group Lasso

- Dans certaines applications, les variables explicatives appartiennent à des groupes de variables prédéfinis.
- On est alors amené à "shrinker" ou à sélectionner les variables par groupe.

Exemple : variables qualitatives

- 2 variables explicatives qualitatives X_1 et X_2 et une variable explicative continu X_3.
- Le modèle logistique s'écrit

$$\begin{aligned}\text{logit}\, p_\beta(x) =& \beta_0 + \beta_1 \mathbf{1}_{x_1=A} + \beta_2 \mathbf{1}_{x_1=B} + \beta_3 \mathbf{1}_{x_1=C} \\ &+ \beta_4 \mathbf{1}_{x_2=D} + \beta_5 \mathbf{1}_{x_2=E} + \beta_6 \mathbf{1}_{x_2=F} + \beta_7 \mathbf{1}_{x_2=G} + \beta_8 X_3\end{aligned}$$

  muni des contraintes $\beta_1 = \beta_4 = 0$.
- 3 groupes : $\mathbf{X}_1 = (\mathbf{1}_{x_1=B}, \mathbf{1}_{x_1=C})$, $\mathbf{X}_2 = (\mathbf{1}_{x_2=E}, \mathbf{1}_{x_2=F}, \mathbf{1}_{x_2=G})$ et $\mathbf{X}_3 = x_3$.

Group Lasso

- Dans certaines applications, les variables explicatives appartiennent à des groupes de variables prédéfinis.
- On est alors amené à "shrinker" ou à sélectionner les variables par groupe.

Exemple : variables qualitatives

- 2 variables explicatives qualitatives X_1 et X_2 et une variable explicative continu X_3.
- Le modèle logistique s'écrit

$$\begin{aligned}\text{logit}\, p_\beta(x) =& \beta_0 + \beta_1 \mathbf{1}_{x_1=A} + \beta_2 \mathbf{1}_{x_1=B} + \beta_3 \mathbf{1}_{x_1=C} \\ &+ \beta_4 \mathbf{1}_{x_2=D} + \beta_5 \mathbf{1}_{x_2=E} + \beta_6 \mathbf{1}_{x_2=F} + \beta_7 \mathbf{1}_{x_2=G} + \beta_8 X_3\end{aligned}$$

  muni des contraintes $\beta_1 = \beta_4 = 0$.
- 3 groupes : $\mathbf{X}_1 = (\mathbf{1}_{x_1=B}, \mathbf{1}_{x_1=C})$, $\mathbf{X}_2 = (\mathbf{1}_{x_2=E}, \mathbf{1}_{x_2=F}, \mathbf{1}_{x_2=G})$ et $\mathbf{X}_3 = x_3$.

Group Lasso

- Dans certaines applications, les variables explicatives appartiennent à des groupes de variables prédéfinis.
- On est alors amené à "shrinker" ou à sélectionner les variables par groupe.

Exemple : variables qualitatives

- 2 variables explicatives qualitatives X_1 et X_2 et une variable explicative continu X_3.
- Le modèle logistique s'écrit

$$\begin{aligned}\text{logit}\, p_\beta(x) =& \beta_0 + \beta_1 \mathbf{1}_{x_1=A} + \beta_2 \mathbf{1}_{x_1=B} + \beta_3 \mathbf{1}_{x_1=C} \\ &+ \beta_4 \mathbf{1}_{x_2=D} + \beta_5 \mathbf{1}_{x_2=E} + \beta_6 \mathbf{1}_{x_2=F} + \beta_7 \mathbf{1}_{x_2=G} + \beta_8 X_3\end{aligned}$$

  muni des contraintes $\beta_1 = \beta_4 = 0$.
- 3 groupes : $\mathbf{X}_1 = (1_{x_1=B}, 1_{x_1=C})$, $\mathbf{X}_2 = (1_{x_2=E}, 1_{x_2=F}, 1_{x_2=G})$ et $\mathbf{X}_3 = x_3$.

Group Lasso

- Dans certaines applications, les variables explicatives appartiennent à des groupes de variables prédéfinis.
- On est alors amené à "shrinker" ou à sélectionner les variables par groupe.

Exemple : variables qualitatives

- 2 variables explicatives qualitatives X_1 et X_2 et une variable explicative continu X_3.
- Le modèle logistique s'écrit

$$\begin{aligned}\text{logit}\, p_\beta(x) =& \beta_0 + \beta_1 \mathbf{1}_{x_1=A} + \beta_2 \mathbf{1}_{x_1=B} + \beta_3 \mathbf{1}_{x_1=C} \\ &+ \beta_4 \mathbf{1}_{x_2=D} + \beta_5 \mathbf{1}_{x_2=E} + \beta_6 \mathbf{1}_{x_2=F} + \beta_7 \mathbf{1}_{x_2=G} + \beta_8 X_3\end{aligned}$$

  muni des contraintes $\beta_1 = \beta_4 = 0$.
- 3 groupes : $\mathbf{X}_1 = (\mathbf{1}_{x_1=B}, \mathbf{1}_{x_1=C})$, $\mathbf{X}_2 = (\mathbf{1}_{x_2=E}, \mathbf{1}_{x_2=F}, \mathbf{1}_{x_2=G})$ et $\mathbf{X}_3 = x_3$.

Définition

En présence de d variables réparties en L groupes $\mathbf{X}_1, \ldots, \mathbf{X}_L$ de cardinal $d_1, \ldots, d_L$. On note $\tilde{\beta}_\ell, \ell = 1, \ldots, L$ le vecteur des coefficients associé au groupe $\mathbf{X}_\ell$. Les estimateurs group-lasso s'obtiennent en maximisant le critère

$$\sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} - \lambda \sum_{\ell=1}^{L} \sqrt{d_\ell} \|\tilde{\beta}_\ell\|_2$$

Remarque

Puisque $\|\tilde{\beta}_\ell\|_2 = 0$ ssi $\tilde{\beta}_{\ell 1} = \ldots = \tilde{\beta}_{\ell d_\ell} = 0$, cette procédure encourage la mise à zéro des coefficients d'un même groupe.

Définition

En présence de d variables réparties en L groupes $\mathbf{X}_1, \ldots, \mathbf{X}_L$ de cardinal $d_1, \ldots, d_L$. On note $\tilde{\beta}_\ell, \ell = 1, \ldots, L$ le vecteur des coefficients associé au groupe $\mathbf{X}_\ell$. Les estimateurs group-lasso s'obtiennent en maximisant le critère

$$\sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} - \lambda \sum_{\ell=1}^{L} \sqrt{d_\ell} \|\tilde{\beta}_\ell\|_2$$

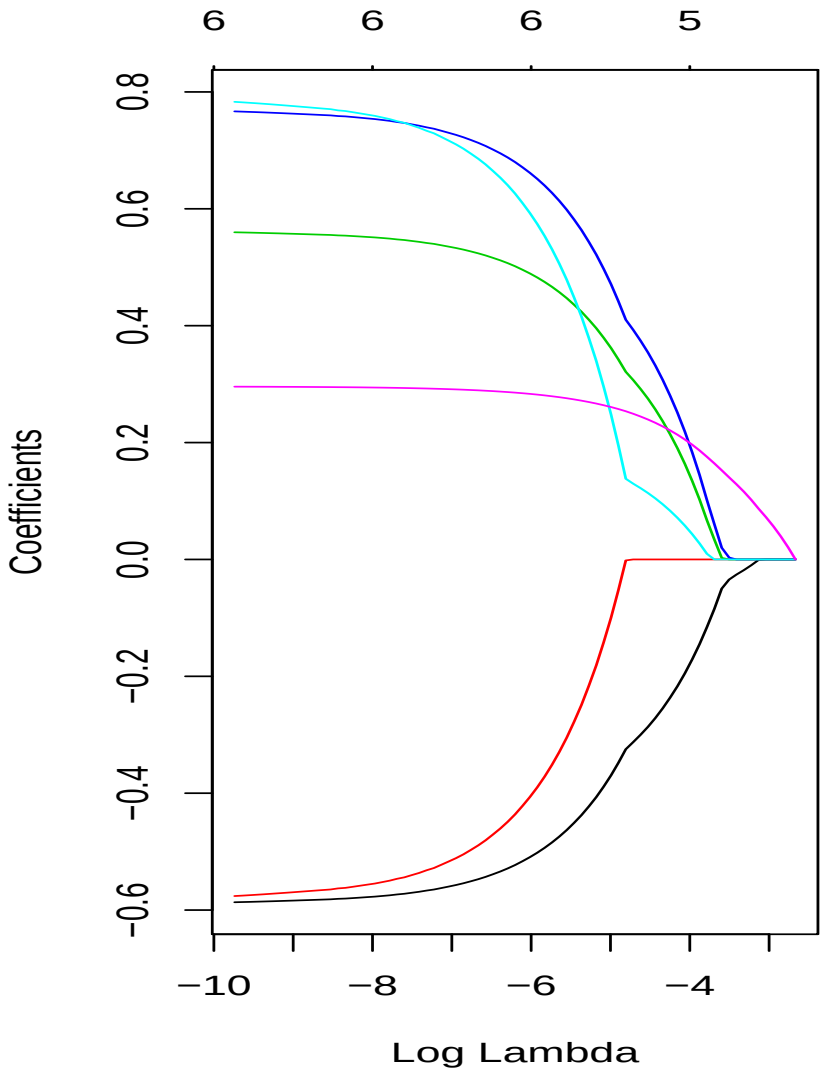
Remarque

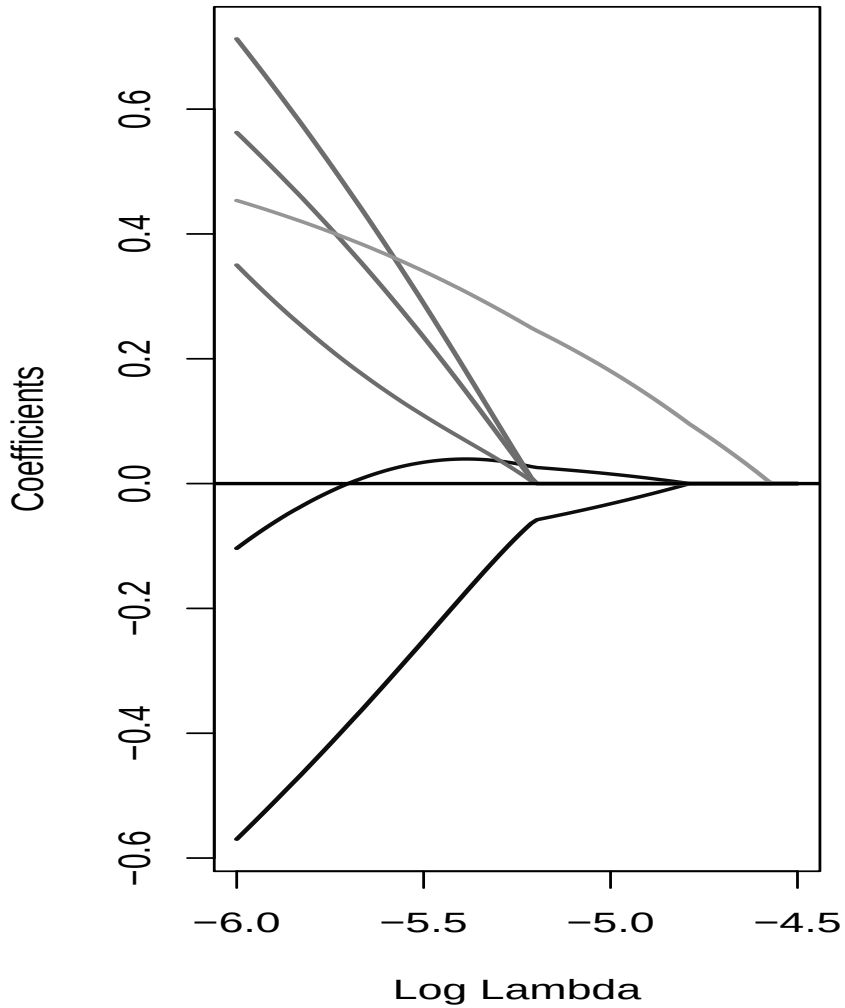
Puisque $\|\tilde{\beta}_\ell\|_2 = 0$ ssi $\tilde{\beta}_{\ell 1} = \ldots = \tilde{\beta}_{\ell d_\ell} = 0$, cette procédure encourage la mise à zéro des coefficients d'un même groupe.

Le coin R

- La fonction gglasso du package gglasso permet de faire du groupe lasso sur R.

```
> summary(donnees)
 X1     X2          X_3                  Y
 A:60   E:40   Min.   :0.002219   Min.   :-1.00
 B:90   F:60   1st Qu.:0.252642   1st Qu.:-1.00
 C:50   G:55   Median :0.505703   Median : 1.00
        H:45   Mean   :0.508092   Mean   : 0.05
               3rd Qu.:0.745967   3rd Qu.: 1.00
               Max.   :0.995240   Max.   : 1.00
> D <- model.matrix(Y~.,data=donnees)[,-1]
> model <- glmnet(D,Y,alpha=1)
> plot(model,label=TRUE,xvar="lambda",lwd=2)
> groupe <- c(1,1,2,2,2,3)
> library(glasso)
> model1 <- gglasso(D,Y,group=groupe,loss="logit",
                    lambda=seq(0.001,0.04,length=100))
> plot(model1)
```





Les coefficients s'annulent par groupe lorsque λ augmente (graphe de droite).

Elastic net

- [Zou and Hastie, 2005] ont proposer de combiner les approches ridge et lasso en proposant une pénalité (appelée elastic net) de la forme

$$\lambda \sum_{j=1}^{d} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

où $\alpha \in [0, 1]$.

- Le paramètre α définit le compromis ridge/lasso :
  - $\alpha = 0 \Longrightarrow$ Lasso ;
  - $\alpha = 1 \Longrightarrow$ Ridge ;
  - Ce paramètre correspond (évidemment) à l'argument `alpha` de la fonction glmnet.
- Avantage : on a plus de flexibilité car la pénalité elastic net propose une gamme de modèles beaucoup plus large que lasso et ridge ;
- Inconvénient : en plus du λ il faut aussi sélectionner le α !

Elastic net

- [Zou and Hastie, 2005] ont proposer de combiner les approches ridge et lasso en proposant une pénalité (appelée elastic net) de la forme

$$\lambda \sum_{j=1}^{d} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

où $\alpha \in [0,1]$.

- Le paramètre α définit le compromis ridge/lasso :
  - $\alpha = 0 \Longrightarrow$ Lasso ;
  - $\alpha = 1 \Longrightarrow$ Ridge ;
  - Ce paramètre correspond (évidemment) à l'argument `alpha` de la fonction glmnet.
- Avantage : on a plus de flexibilité car la pénalité elastic net propose une gamme de modèles beaucoup plus large que lasso et ridge ;
- Inconvénient : en plus du λ il faut aussi sélectionner le α !

Elastic net

- [Zou and Hastie, 2005] ont proposer de combiner les approches ridge et lasso en proposant une pénalité (appelée elastic net) de la forme

$$\lambda \sum_{j=1}^{d} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

où $\alpha \in [0, 1]$.

- Le paramètre α définit le compromis ridge/lasso :
  - $\alpha = 0 \Longrightarrow$ Lasso ;
  - $\alpha = 1 \Longrightarrow$ Ridge ;
  - Ce paramètre correspond (évidemment) à l'argument `alpha` de la fonction glmnet.
- Avantage : on a plus de flexibilité car la pénalité elastic net propose une gamme de modèles beaucoup plus large que lasso et ridge ;
- Inconvénient : en plus du λ il faut aussi sélectionner le α !

Elastic net

- [Zou and Hastie, 2005] ont proposer de combiner les approches ridge et lasso en proposant une pénalité (appelée elastic net) de la forme

$$\lambda \sum_{j=1}^{d} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

où $\alpha \in [0, 1]$.

- Le paramètre α définit le compromis ridge/lasso :
  - $\alpha = 0 \Longrightarrow$ Lasso ;
  - $\alpha = 1 \Longrightarrow$ Ridge ;
  - Ce paramètre correspond (évidemment) à l'argument `alpha` de la fonction glmnet.
- Avantage : on a plus de flexibilité car la pénalité elastic net propose une gamme de modèles beaucoup plus large que lasso et ridge ;
- Inconvénient : en plus du λ il faut aussi sélectionner le α !

Elastic net

- [Zou and Hastie, 2005] ont proposer de combiner les approches ridge et lasso en proposant une pénalité (appelée elastic net) de la forme

$$\lambda \sum_{j=1}^{d} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

où $\alpha \in [0, 1]$.

- Le paramètre α définit le compromis ridge/lasso :
  - $\alpha = 0 \Longrightarrow$ Lasso ;
  - $\alpha = 1 \Longrightarrow$ Ridge ;
  - Ce paramètre correspond (évidemment) à l'argument `alpha` de la fonction glmnet.
- Avantage : on a plus de flexibilité car la pénalité elastic net propose une gamme de modèles beaucoup plus large que lasso et ridge ;
- Inconvénient : en plus du λ il faut aussi sélectionner le α !

Références I


Bühlmann, P. and van de Geer, S. (2011).
Statistics for high-dimensional data.
Springer.


Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society, Series B, 58 :267–288.


Zou, H. and Hastie, T. (2005).
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society, Series B, 67 :301–320.

Septième partie VII

Introduction au scoring

Objectifs du chapitre

- Présenter une des méthodes phares dans les études de statistique appliquée : le scoring.
- Nombreux domaines d'applications (banque, assurance, plus généralement : risque et marketing).
- Un score permet d'attribuer à chaque individu une note représentant (ou estimant) la probabilité de cliblage d'un évènement :
  - remboursement d'un crédit
  - souscription à une assurance vie
  - consommation d'un produit
  - changement d'opérateur.

Objectifs du chapitre

- Présenter une des méthodes phares dans les études de statistique appliquée : le scoring.
- Nombreux domaines d'applications (banque, assurance, plus généralement : risque et marketing).
- Un score permet d'attribuer à chaque individu une note représentant (ou estimant) la probabilité de cliblage d'un évènement :
  - remboursement d'un crédit
  - souscription à une assurance vie
  - consommation d'un produit
  - changement d'opérateur.

Objectifs du chapitre

- Présenter une des méthodes phares dans les études de statistique appliquée : le scoring.
- Nombreux domaines d'applications (banque, assurance, plus généralement : risque et marketing).
- Un score permet d'attribuer à chaque individu une note représentant (ou estimant) la probabilité de cliblage d'un évènement :
  - remboursement d'un crédit
  - souscription à une assurance vie
  - consommation d'un produit
  - changement d'opérateur.

Etapes de construction d'un score

1. Construction de la base d'étude
   - Définition de l'évènement à étudier
   - Définition de la population éligible
   - Définition de la période d'étude
   - Construction de variables explicatives
2. Modélisation
   - Construction de plusieurs modèles
   - Choix du "meilleur" modèle
   - Interprétation du modèle choisi
3. Exploitation du score
   - Application du score
   - Suivi de la performance du score - mise à jour du score.

Etapes de construction d'un score

1. Construction de la base d'étude
   - Définition de l'évènement à étudier
   - Définition de la population éligible
   - Définition de la période d'étude
   - Construction de variables explicatives
2. Modélisation
   - Construction de plusieurs modèles
   - Choix du "meilleur" modèle
   - Interprétation du modèle choisi
3. Exploitation du score
   - Application du score
   - Suivi de la performance du score - mise à jour du score.

Etapes de construction d'un score

1. Construction de la base d'étude
   - Définition de l'évènement à étudier
   - Définition de la population éligible
   - Définition de la période d'étude
   - Construction de variables explicatives
2. Modélisation
   - Construction de plusieurs modèles
   - Choix du "meilleur" modèle
   - Interprétation du modèle choisi
3. Exploitation du score
   - Application du score
   - Suivi de la performance du score - mise à jour du score.

- **Objectif :** construire une base de données propre.
- Etape la plus longue d'un projet de scoring.
- Etape importante car les choix effectués auront une influence sur la performance du futur modèle.

- **Objectif :** construire une base de données propre.
- Etape la plus longue d'un projet de scoring.
- Etape importante car les choix effectués auront une influence sur la performance du futur modèle.

Evènement à étudier

- L'évènement à étudier n'est pas forcément présent dans la base de données initiales. Il doit être défini en fonction de l'objectif de l'étude.

Exemple : crédit scoring

- L'objectif est de prévenir le risque d'impayés pour les clients à qui on accorde un crédit.
- La base de données renseigne uniquement sur le nombre d'impayés pour les clients présents dans l'historique.
- On construira une variable Y qui vaut 1 si un client a connu plus de K impayés, 0 sinon (la valeur de K dépend des objectifs du moment).

- Cette étape permet la construction de la variable à expliquer.
- En général, cette variable à expliquer est binaire et permet de comparer (discriminer) 2 sous-populations (individus ayant réalisé l'évènement qui auront la valeur "1", ceux ne l'ayant pas réalisé auront la valeur "0").

Evènement à étudier

- L'évènement à étudier n'est pas forcément présent dans la base de données initiales. Il doit être défini en fonction de l'objectif de l'étude.

Exemple : crédit scoring

- L'objectif est de prévenir le risque d'impayés pour les clients à qui on accorde un crédit.
- La base de données renseigne uniquement sur le nombre d'impayés pour les clients présents dans l'historique.
- On construira une variable Y qui vaut 1 si un client a connu plus de K impayés, 0 sinon (la valeur de K dépend des objectifs du moment).

- Cette étape permet la construction de la variable à expliquer.
- En général, cette variable à expliquer est binaire et permet de comparer (discriminer) 2 sous-populations (individus ayant réalisé l'évènement qui auront la valeur "1", ceux ne l'ayant pas réalisé auront la valeur "0").

Population éligible

- La définition de la population éligible permet d'identifier l'ensemble individus à inclure dans l'étude.
- Elle est constituée de clients dont les caractéristiques doivent être identiques à celles des clients sur lesquels le score sera appliqué.

Exemples de critères à utiliser

1. Liés à l'évènement à étudier : intégrer uniquement les individus pour lesquels l'évènement peut se réaliser.
2. Liés à des choix stratégiques (experts métiers) : âge, ancienneté...

Population éligible

- La définition de la population éligible permet d'identifier l'ensemble individus à inclure dans l'étude.
- Elle est constituée de clients dont les caractéristiques doivent être identiques à celles des clients sur lesquels le score sera appliqué.

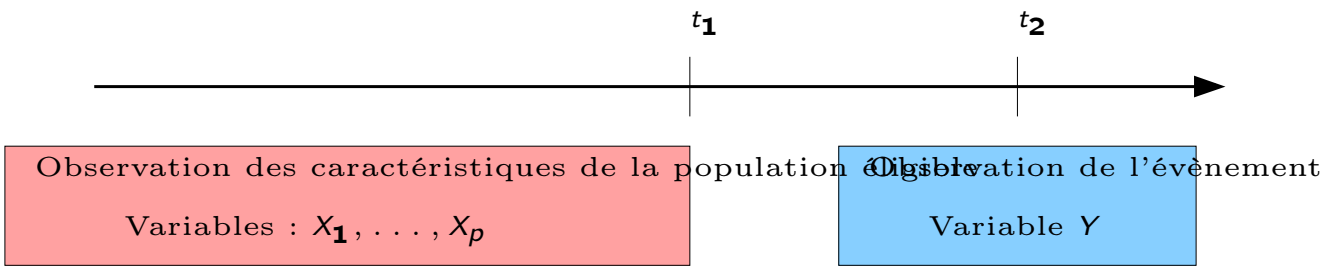
Exemples de critères à utiliser

1. Liés à l'évènement à étudier : intégrer uniquement les individus pour lesquels l'évènement peut se réaliser.
2. Liés à des choix stratégiques (experts métiers) : âge, ancienneté...

Période d'étude

La période d'étude permet de définir le moment où l'évènement (Y) est observé et le moment où le comportement du client est analysé.

- Elle est définie par deux instants :
  1. t_1 : date de référence. L'évènement n'a pas encore été observé sur la population éligible.
  2. t_2 : date à laquelle on observe l'évènement.



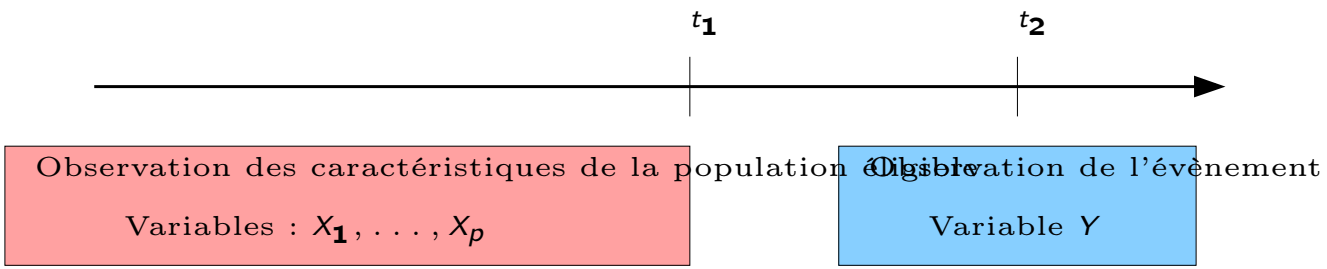
Remarque

Ce processus peut-être réitéré sur autant de périodes que nécessaire afin d'obtenir suffisamment de données.

Période d'étude

La période d'étude permet de définir le moment où l'évènement (Y) est observé et le moment où le comportement du client est analysé.

- Elle est définie par deux instants :
  1. t_1 : date de référence. L'évènement n'a pas encore été observé sur la population éligible.
  2. t_2 : date à laquelle on observe l'évènement.



Remarque

Ce processus peut-être réitéré sur autant de périodes que nécessaire afin d'obtenir suffisamment de données.

- A l'issue de cette phase, on dispose d'une variable à expliquer et d'un certain nombre de variables explicatives.
- L'objectif est bien entendu de "trouver" une fonction m telle que

$$\text{variable à expliquer} \approx m(\text{variables explicatives}).$$

- Nécessité de poser un modèle statistique puis d'estimer m dans le modèle considéré.
- Il est souvent nécessaire de "travailler" au préalable sur les variables explicatives.

- A l'issue de cette phase, on dispose d'une variable à expliquer et d'un certain nombre de variables explicatives.
- L'objectif est bien entendu de "trouver" une fonction m telle que

$$\text{variable à expliquer} \approx m(\text{variables explicatives}).$$

- Nécessité de poser un modèle statistique puis d'estimer m dans le modèle considéré.
- Il est souvent nécessaire de "travailler" au préalable sur les variables explicatives.

- A l'issue de cette phase, on dispose d'une variable à expliquer et d'un certain nombre de variables explicatives.
- L'objectif est bien entendu de "trouver" une fonction m telle que

$$\text{variable à expliquer} \approx m(\text{variables explicatives}).$$

- Nécessité de poser un modèle statistique puis d'estimer m dans le modèle considéré.
- Il est souvent nécessaire de "travailler" au préalable sur les variables explicatives.

- A l'issue de cette phase, on dispose d'une variable à expliquer et d'un certain nombre de variables explicatives.
- L'objectif est bien entendu de "trouver" une fonction m telle que

$$\text{variable à expliquer} \approx m(\text{variables explicatives}).$$

- Nécessité de poser un modèle statistique puis d'estimer m dans le modèle considéré.
- Il est souvent nécessaire de "travailler" au préalable sur les variables explicatives.

Variables explicatives

- Elles correspondent à tous les indicateurs mesurables pouvant potentiellement expliquer le phénomène considéré.
- Il est généralement possible d'ajuster directement un modèle sur ces variables brutes...
- mais... il est souvent préférable de "travailler" sur ces variables (les étudier, en supprimer, en transformer...) pour pouvoir obtenir des modèles plus performants par la suite.

Variables explicatives

- Elles correspondent à tous les indicateurs mesurables pouvant potentiellement expliquer le phénomène considéré.
- Il est généralement possible d'ajuster directement un modèle sur ces variables brutes...
- mais... il est souvent préférable de "travailler" sur ces variables (les étudier, en supprimer, en transformer...) pour pouvoir obtenir des modèles plus performants par la suite.

Variables explicatives

- Elles correspondent à tous les indicateurs mesurables pouvant potentiellement expliquer le phénomène considéré.
- Il est généralement possible d'ajuster directement un modèle sur ces variables brutes...
- mais... il est souvent préférable de "travailler" sur ces variables (les étudier, en supprimer, en transformer...) pour pouvoir obtenir des modèles plus performants par la suite.

Construction de variables

- En plus des variables brutes, il est indispensable de construire de nouveaux indicateurs (nouvelles variables) :
  - Moyenne / médiane de plusieurs variables
  - Ratio / taux d'accroissements
  - Evolutions entre plusieurs dates
  - Croisements de variables
  - ...

Exemple

- On souhaite expliquer la fidélité de clients vis à vis d'un produit ou d'un forfait.
- On dispose de la date de souscription de ce forfait.
- Calculer "l'âge de souscription" à la date de référence du client.

Construction de variables

- En plus des variables brutes, il est indispensable de construire de nouveaux indicateurs (nouvelles variables) :
  - Moyenne / médiane de plusieurs variables
  - Ratio / taux d'accroissements
  - Evolutions entre plusieurs dates
  - Croisements de variables
  - ...

Exemple

- On souhaite expliquer la fidélité de clients vis à vis d'un produit ou d'un forfait.
- On dispose de la date de souscription de ce forfait.
- Calculer "l'âge de souscription" à la date de référence du client.

Construction de variables

- En plus des variables brutes, il est indispensable de construire de nouveaux indicateurs (nouvelles variables) :
  - Moyenne / médiane de plusieurs variables
  - Ratio / taux d'accroissements
  - Evolutions entre plusieurs dates
  - Croisements de variables
  - ...

Exemple

- On souhaite expliquer la fidélité de clients vis à vis d'un produit ou d'un forfait.
- On dispose de la date de souscription de ce forfait.
- Calculer "l'âge de souscription" à la date de référence du client.

- Chaque variable doit ensuite être "fiabilisée" à l'aide de statistiques descriptives :
  - Calcul d'indicateurs de tendance centrale - de dispersion.
  - Analyses factorielles (ACP/ACM).

- Cette analyse doit permettre de détecter et de traiter les points suivants :
  - variables "inutiles"/colinéarité entre variables explicatives
  - valeurs manquantes/aberrantes/extrêmes
  - modalités à faibles effectifs
  - incohérence...

- Chaque variable doit ensuite être "fiabilisée" à l'aide de statistiques descriptives :
  - Calcul d'indicateurs de tendance centrale - de dispersion.
  - Analyses factorielles (ACP/ACM).

- Cette analyse doit permettre de détecter et de traiter les points suivants :
  - variables "inutiles"/colinéarité entre variables explicatives
  - valeurs manquantes/aberrantes/extrêmes
  - modalités à faibles effectifs
  - incohérence...

Discrétisation

- Selon le modèle utilisé par la suite, les variables explicatives quantitatives peuvent être discrétisées (regroupées en classe).
- Avantages :
  - Prise en compte d'effets non linéaires (intéressant pour la régression logistique).
  - Permet de gérer facilement les valeurs manquantes.
- Incovénients
  - Augmentatation du nombre de paramètres à estimer (pour les modèles paramétriques) et donc de la variance des estimateurs.
  - Il n'existe pas de règles universelles optimales de discrétisation (chacun fait sa cuisine).

Discrétisation

- Selon le modèle utilisé par la suite, les variables explicatives quantitatives peuvent être discrétisées (regroupées en classe).
- Avantages :
  - Prise en compte d'effets non linéaires (intéressant pour la régression logistique).
  - Permet de gérer facilement les valeurs manquantes.
- Incovénients
  - Augmentatation du nombre de paramètres à estimer (pour les modèles paramétriques) et donc de la variance des estimateurs.
  - Il n'existe pas de règles universelles optimales de discrétisation (chacun fait sa cuisine).

Discrétisation

- Selon le modèle utilisé par la suite, les variables explicatives quantitatives peuvent être discrétisées (regroupées en classe).
- Avantages :
  - Prise en compte d'effets non linéaires (intéressant pour la régression logistique).
  - Permet de gérer facilement les valeurs manquantes.
- Incovénients
  - Augmentatation du nombre de paramètres à estimer (pour les modèles paramétriques) et donc de la variance des estimateurs.
  - Il n'existe pas de règles universelles optimales de discrétisation (chacun fait sa cuisine).

Les données

- A l'issue de cette (longue) étape, le statisticien dispose (enfin) de sa base de données.
- Elle est consitutée de l'observation de p variables explicatives $\mathbf{X_1}, \ldots, \mathbf{X_p}$ et d'une variable à expliquer Y mesurées sur n individus.
- On peut résumer ces données à l'aide d'une matrice

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} & y_1 \\ \vdots & & & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} & y_n \end{pmatrix}$$

Le problème

Il consiste à expliquer Y par $\mathbf{X_1}, \ldots, \mathbf{X_p}$ ou encore à prédire Y à partir de $\mathbf{X_1}, \ldots, \mathbf{X_p}$ à l'aide des données ci-dessus.

Les données

- A l'issue de cette (longue) étape, le statisticien dispose (enfin) de sa base de données.
- Elle est consitutée de l'observation de p variables explicatives $\mathbf{X_1}, \ldots, \mathbf{X_p}$ et d'une variable à expliquer Y mesurées sur n individus.
- On peut résumer ces données à l'aide d'une matrice

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} & y_1 \\ \vdots & & & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} & y_n \end{pmatrix}$$

Le problème

Il consiste à expliquer Y par $\mathbf{X_1}, \ldots, \mathbf{X_p}$ ou encore à prédire Y à partir de $\mathbf{X_1}, \ldots, \mathbf{X_p}$ à l'aide des données ci-dessus.

Les données

- A l'issue de cette (longue) étape, le statisticien dispose (enfin) de sa base de données.
- Elle est consitutée de l'observation de p variables explicatives $\mathbf{X_1}, \ldots, \mathbf{X_p}$ et d'une variable à expliquer Y mesurées sur n individus.
- On peut résumer ces données à l'aide d'une matrice

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} & y_1 \\ \vdots & & & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} & y_n \end{pmatrix}$$

Le problème

Il consiste à expliquer Y par $\mathbf{X_1}, \ldots, \mathbf{X_P}$ ou encore à prédire Y à partir de $\mathbf{X_1}, \ldots, \mathbf{X_P}$ à l'aide des données ci-dessus.

Risque bancaire

- Un individu souhaite réaliser crédit bancaire.
- Le banquier demande un certain nombre d'information (âge, sexe, CSP, revenus...).
- Il les saisit dans sa machine...
- qui lui renvoie une réponse (oui ou non).

Question

Que fait la machine ?

Risque bancaire

- Un individu souhaite réaliser crédit bancaire.
- Le banquier demande un certain nombre d'information (âge, sexe, CSP, revenus...).
- Il les saisit dans sa machine...
- qui lui renvoie une réponse (oui ou non).

Question

Que fait la machine ?

Risque bancaire

- Un individu souhaite réaliser crédit bancaire.
- Le banquier demande un certain nombre d'information (âge, sexe, CSP, revenus...).
- Il les saisit dans sa machine...
- qui lui renvoie une réponse (oui ou non).

Question

Que fait la machine ?

Risque bancaire

- Un individu souhaite réaliser crédit bancaire.
- Le banquier demande un certain nombre d'information (âge, sexe, CSP, revenus...).
- Il les saisit dans sa machine...
- qui lui renvoie une réponse (oui ou non).

Question

Que fait la machine ?

Marketing

- Une entreprise souhaite booster les ventes d'un produit auprès de ces clients.
- Elle souhaite envoyer une promotion à ses clients les plus appétents à ce produit.

Question

Comment les sélectionner ?

Marketing

- Une entreprise souhaite booster les ventes d'un produit auprès de ces clients.
- Elle souhaite envoyer une promotion à ses clients les plus appétents à ce produit.

Question

Comment les sélectionner ?

Churn

- Utilisé par les services clients en téléphonie.
- L'objectif est de tenter d'identifier les clients susceptibles de partir vers la concurrence afin d'essayer de les retenir (en leur proposant une offre par exemple).
- Appeler tous les clients régulièrement à un coût, il est donc nécessaire de cibler les bons clients au bon moment.

Question

Comment identifier ces "mauvais" clients ?

Churn

- Utilisé par les services clients en téléphonie.
- L'objectif est de tenter d'identifier les clients susceptibles de partir vers la concurrence afin d'essayer de les retenir (en leur proposant une offre par exemple).
- Appeler tous les clients régulièrement à un coût, il est donc nécessaire de cibler les bons clients au bon moment.

Question

Comment identifier ces "mauvais" clients ?

Churn

- Utilisé par les services clients en téléphonie.
- L'objectif est de tenter d'identifier les clients susceptibles de partir vers la concurrence afin d'essayer de les retenir (en leur proposant une offre par exemple).
- Appeler tous les clients régulièrement à un coût, il est donc nécessaire de cibler les bons clients au bon moment.

Question

Comment identifier ces "mauvais" clients ?

Score

- Pour ces 3 problèmes (et pour bien d'autre encore), il s'agit de faire un choix entre deux issues :
  - acceptation ou rejet du crédit.
  - envoi ou non de l'offre au client.
  - proposer un renouvellement (fidéliser) au client.

Une solution

Modéliser ces deux issues à l'aide d'une variable binaire Y que l'on cherche à prédire par $\hat{Y}$.

Score

- Pour ces 3 problèmes (et pour bien d'autre encore), il s'agit de faire un choix entre deux issues :
  - acceptation ou rejet du crédit.
  - envoi ou non de l'offre au client.
  - proposer un renouvellement (fidéliser) au client.

Une solution

Modéliser ces deux issues à l'aide d'une variable binaire Y que l'on cherche à prédire par $\hat{Y}$.

Score

- Pour ces 3 problèmes (et pour bien d'autre encore), il s'agit de faire un choix entre deux issues :
  - acceptation ou rejet du crédit.
  - envoi ou non de l'offre au client.
  - proposer un renouvellement (fidéliser) au client.

Une solution

Modéliser ces deux issues à l'aide d'une variable binaire Y que l'on cherche à prédire par $\hat{Y}$.

Un début de modélisation

Ces trois exemples peuvent être traités à l'aide d'un modèle de régression comprenant

- une variable à expliquer binaire Y en lien avec la réponse que l'on doit donner :
  - Y vaut 0 si le client est "bon", 1 si il est mauvais.
  - Y vaut 0 si le client a déjà consommé le produit, 1 sinon.
  - Y vaut 0 si le client est fidèle à l'opérateur, 1 si il en a changé.
- p variables explicatives susceptibles d'aider à comprendre Y :
  - âge, revenus...
  - fréquence d'achat, âge, domicile...
  - date du souscription de précédent contrat...

Un début de modélisation

Ces trois exemples peuvent être traités à l'aide d'un modèle de régression comprenant

- une variable à expliquer binaire Y en lien avec la réponse que l'on doit donner :
  - Y vaut 0 si le client est "bon", 1 si il est mauvais.
  - Y vaut 0 si le client a déjà consommé le produit, 1 sinon.
  - Y vaut 0 si le client est fidèle à l'opérateur, 1 si il en a changé.
- p variables explicatives susceptibles d'aider à comprendre Y :
  - âge, revenus...
  - fréquence d'achat, âge, domicile...
  - date du souscription de précédent contrat...

Un début de modélisation

Ces trois exemples peuvent être traités à l'aide d'un modèle de régression comprenant

- une variable à expliquer binaire Y en lien avec la réponse que l'on doit donner :
  - Y vaut 0 si le client est "bon", 1 si il est mauvais.
  - Y vaut 0 si le client a déjà consommé le produit, 1 sinon.
  - Y vaut 0 si le client est fidèle à l'opérateur, 1 si il en a changé.
- p variables explicatives susceptibles d'aider à comprendre Y :
  - âge, revenus...
  - fréquence d'achat, âge, domicile...
  - date du souscription de précédent contrat...

- Ces variables sont mesurées sur n individus. Cela nous amène à considérer un n échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ supposé i.i.d tel que
  - X_i est un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ représentant les variables explicatives du $i$ème individu.
  - Y_i est une variable aléatoire binaire représentant la variable à expliquer du $i$ème individu.

Remarque

- L'indépendance revient à supposer les individus n'ont pas d'influence les uns sur les autres (le fait que i rembourse mal ses crédits n'influence pas les remboursements de j).
- "identiquement distribué" revient à dire que les n individus sont issus d'une "même population" (on ne prend pas des individus des années 1950 et des années 2010 pour expliquer le remboursement de crédit).

- Ces variables sont mesurées sur n individus. Cela nous amène à considérer un n échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ supposé i.i.d tel que
  - X_i est un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ représentant les variables explicatives du $i$ème individu.
  - Y_i est une variable aléatoire binaire représentant la variable à expliquer du $i$ème individu.

Remarque

- L'indépendance revient à supposer les individus n'ont pas d'influence les uns sur les autres (le fait que i rembourse mal ses crédits n'influence pas les remboursements de j).
- "identiquement distribué" revient à dire que les n individus sont issus d'une "même population" (on ne prend pas des individus des années 1950 et des années 2010 pour expliquer le remboursement de crédit).

- Ces variables sont mesurées sur n individus. Cela nous amène à considérer un n échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ supposé i.i.d tel que
  - X_i est un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ représentant les variables explicatives du $i$ème individu.
  - Y_i est une variable aléatoire binaire représentant la variable à expliquer du $i$ème individu.

Remarque

- L'indépendance revient à supposer les individus n'ont pas d'influence les uns sur les autres (le fait que i rembourse mal ses crédits n'influence pas les remboursements de j).
- "identiquement distribué" revient à dire que les n individus sont issus d'une "même population" (on ne prend pas des individus des années 1950 et des années 2010 pour expliquer le remboursement de crédit).

Le modèle logistique

- De nombreux modèles statistiques permettent de traiter ce genre de problème.
- Le modèle de régression logistique par exemple :
  $\mathcal{L}(Y|X = x) = \mathcal{B}(p(x))$ telle que

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

- Les paramètres β_j sont estimés par maximum de vraisemblance à l'aide d'un n-échantillon $(x_1, y_1), \ldots, (x_n, y_n)$.
- On peut faire de la prévision selon

$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{p}(x) \geq 0.5 \\ 0 & \text{sinon.} \end{cases}$$

Le modèle logistique

- De nombreux modèles statistiques permettent de traiter ce genre de problème.
- Le modèle de régression logistique par exemple :
  $\mathcal{L}(Y|X = x) = \mathcal{B}(p(x))$ telle que

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

- Les paramètres β_j sont estimés par maximum de vraisemblance à l'aide d'un n-échantillon $(x_1, y_1), \ldots, (x_n, y_n)$.
- On peut faire de la prévision selon

$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{p}(x) \geq 0.5 \\ 0 & \text{sinon.} \end{cases}$$

L'approche scoring

L'approche proposée ici n'est pas totalement satisfaisante :

- selon les objectifs du moment, une banque peut être plus ou moins clémente pour accorder un crédit.
- au cours d'une campagne publicitaire, on peut être tentés de solliciter des clients pour lesquels on est loin d'être certain qu'ils souscrivent au produit.
- oublier un churn est plus grâve que solliciter un client qui reste fidèle.

- Nécessité de disposer d'un outil plus flexible.
- L'approche scoring consiste à donner une note à chaque individu qui soit en relation avec la variable Y.

L'approche scoring

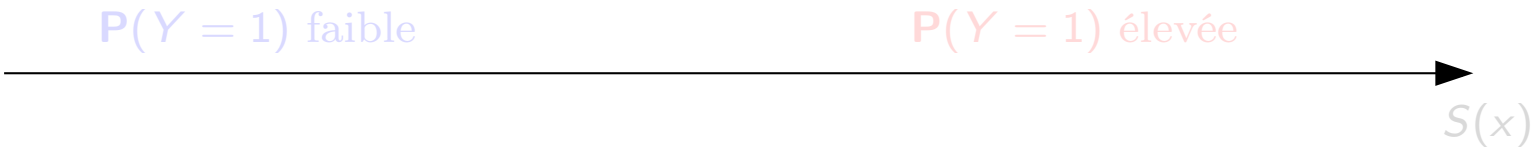
L'approche proposée ici n'est pas totalement satisfaisante :

- selon les objectifs du moment, une banque peut être plus ou moins clémente pour accorder un crédit.
- au cours d'une campagne publicitaire, on peut être tentés de solliciter des clients pour lesquels on est loin d'être certain qu'ils souscrivent au produit.
- oublier un churn est plus grâve que solliciter un client qui reste fidèle.

- Nécessité de disposer d'un outil plus flexible.
- L'approche scoring consiste à donner une note à chaque individu qui soit en relation avec la variable Y.

Score : définition

- Un score est une fonction $S : \mathbb{R}^p \to \mathbb{R}$.
- Etant données n un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ le job du statisticien consiste à construire une fonction $S(x)$ qui permettent d'expliquer Y *au mieux*.



- Une fois le score construit, la décision s'effectue selon la procédure

$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{S}(x) \geq s \\ 0 & \text{sinon,} \end{cases}$$

où s est un seuil choisi par l'utilisateur.

- La construction de scores s'effectue généralement avec les modèles de classification classiques.

Score : définition

- Un score est une fonction $S : \mathbb{R}^p \to \mathbb{R}$.
- Etant données n un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ le job du statisticien consiste à construire une fonction $S(x)$ qui permettent d'expliquer Y *au mieux*.

- Une fois le score construit, la décision s'effectue selon la procédure

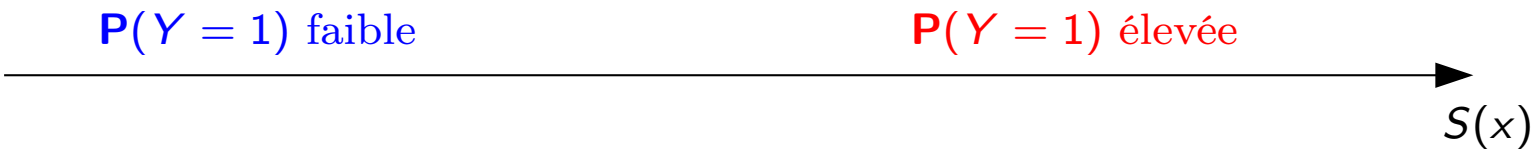
$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{S}(x) \geq s \\ 0 & \text{sinon,} \end{cases}$$

où s est un seuil choisi par l'utilisateur.

- La construction de scores s'effectue généralement avec les modèles de classification classiques.

Score : définition

- Un score est une fonction $S : \mathbb{R}^p \to \mathbb{R}$.
- Etant données n un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ le job du statisticien consiste à construire une fonction $S(x)$ qui permettent d'expliquer Y *au mieux*.



- Une fois le score construit, la décision s'effectue selon la procédure

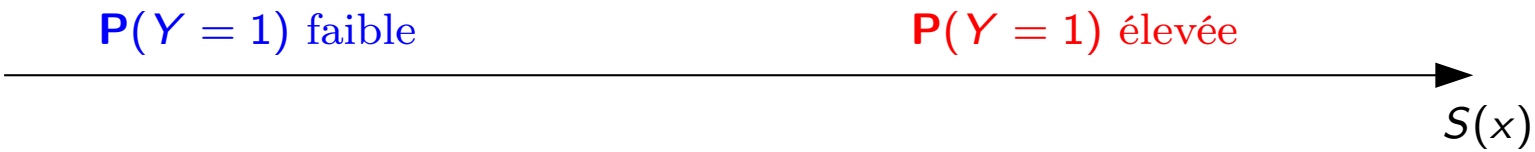
$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{S}(x) \geq s \\ 0 & \text{sinon,} \end{cases}$$

où s est un seuil choisi par l'utilisateur.

- La construction de scores s'effectue généralement avec les modèles de classification classiques.

Score : définition

- Un score est une fonction $S : \mathbb{R}^p \to \mathbb{R}$.
- Etant données n un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ le job du statisticien consiste à construire une fonction $S(x)$ qui permettent d'expliquer Y *au mieux*.



- Une fois le score construit, la décision s'effectue selon la procédure

$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{S}(x) \geq s \\ 0 & \text{sinon,} \end{cases}$$

où s est un seuil choisi par l'utilisateur.

- La construction de scores s'effectue généralement avec les modèles de classification classiques.

- La fonction de score optimale est ainsi définie par

$$S(x) = \mathbf{P}(Y = 1|X = x).$$

- Cette fonction est bien entendue inconnue et le problème du statisticien est d'estimer S à l'aide d'un n-échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- De nombreux modèles de discrimination permettent d'estimer $\mathbf{P}(Y = 1|X = x)$. L'estimation de S sera généralement basée sur les méthodes d'analyse discriminante (les cours de scoring et de discrimination sont fortement liés).

Remarque

- La valeur de la note $S(x)$ n'a pas de réelle importance en scoring. L'important est la manière dont le score va classer des individus $x_1, \ldots, x_n$.
- Il n'est pas forcément nécessaire d'estimer S. On peut se contenter d'une transformation bijective de S.

- La fonction de score optimale est ainsi définie par

$$S(x) = \mathbf{P}(Y = 1|X = x).$$

- Cette fonction est bien entendue inconnue et le problème du statisticien est d'estimer S à l'aide d'un n-échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- De nombreux modèles de discrimination permettent d'estimer $\mathbf{P}(Y = 1|X = x)$. L'estimation de S sera généralement basée sur les méthodes d'analyse discriminante (les cours de scoring et de discrimination sont fortement liés).

Remarque

- La valeur de la note $S(x)$ n'a pas de réelle importance en scoring. L'important est la manière dont le score va classer des individus $x_1, \ldots, x_n$.
- Il n'est pas forcément nécessaire d'estimer S. On peut se contenter d'une transformation bijective de S.

- La fonction de score optimale est ainsi définie par

$$S(x) = \mathbf{P}(Y = 1|X = x).$$

- Cette fonction est bien entendue inconnue et le problème du statisticien est d'estimer S à l'aide d'un n-échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- De nombreux modèles de discrimination permettent d'estimer $\mathbf{P}(Y = 1|X = x)$. L'estimation de S sera généralement basée sur les méthodes d'analyse discriminante (les cours de scoring et de discrimination sont fortement liés).

Remarque

- La valeur de la note $S(x)$ n'a pas de réelle importance en scoring. L'important est la manière dont le score va classer des individus $x_1, \ldots, x_n$.
- Il n'est pas forcément nécessaire d'estimer S. On peut se contenter d'une transformation bijective de S.

Lien score/règle de prévision

- Etant donné un score S, on peut déduire une règle de prévision en fixant un seuil s (la réciproque n'est pas vraie) :

$$g_s(x) = \begin{cases} 1 & \text{si } S(x) \geq s \\ 0 & \text{sinon.} \end{cases}$$

- Cette règle définit la table de confusion

	$g_s(X) = 0$	$g_s(X) = 1$
$Y = 0$	OK	E_1
$Y = 1$	E_2	OK

- Pour chaque seuil s, on distingue deux types d'erreur

$$\alpha(s) = \mathbf{P}(g_s(X) = 1 | Y = 0) = \mathbf{P}(S(X) \geq s | Y = 0)$$

et

$$\beta(s) = \mathbf{P}(g_s(X) = 0 | Y = 1) = \mathbf{P}(S(X) < s | Y = 1).$$

Lien score/règle de prévision

- Etant donné un score S, on peut déduire une règle de prévision en fixant un seuil s (la réciproque n'est pas vraie) :

$$g_s(x) = \begin{cases} 1 & \text{si } S(x) \geq s \\ 0 & \text{sinon.} \end{cases}$$

- Cette règle définit la table de confusion

	$g_s(X) = 0$	$g_s(X) = 1$
$Y = 0$	OK	E_1
$Y = 1$	E_2	OK

- Pour chaque seuil s, on distingue deux types d'erreur

$$\alpha(s) = \mathbf{P}(g_s(X) = 1 | Y = 0) = \mathbf{P}(S(X) \geq s | Y = 0)$$

et

$$\beta(s) = \mathbf{P}(g_s(X) = 0 | Y = 1) = \mathbf{P}(S(X) < s | Y = 1).$$

Score parfait et score aléatoire

On définit également

- Spécificité : $sp(s) = \mathbf{P}(S(X) < s | Y = 0) = 1 - \alpha(s)$
- Sensibilité : $se(s) = \mathbf{P}(S(X) \geq s | Y = 1) = 1 - \beta(s)$

Définition

- *Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que*

$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad et \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

- *Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.*

Performance d'un score

Elle se mesure généralement en visualisant les erreurs $\alpha(s)$ et $\beta(s)$ et/ou la spécificité et la sensibilité pour tous les seuils s.

Score parfait et score aléatoire

On définit également

- Spécificité : $sp(s) = \mathbf{P}(S(X) < s | Y = 0) = 1 - \alpha(s)$
- Sensibilité : $se(s) = \mathbf{P}(S(X) \geq s | Y = 1) = 1 - \beta(s)$

Définition

- *Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que*

$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad et \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

- *Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.*

Performance d'un score

Elle se mesure généralement en visualisant les erreurs $\alpha(s)$ et $\beta(s)$ et/ou la spécificité et la sensibilité pour tous les seuils s.

Score parfait et score aléatoire

On définit également

- Spécificité : $sp(s) = \mathbf{P}(S(X) < s | Y = 0) = 1 - \alpha(s)$
- Sensibilité : $se(s) = \mathbf{P}(S(X) \geq s | Y = 1) = 1 - \beta(s)$

Définition

- *Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que*

$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad et \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

- *Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.*

Performance d'un score

Elle se mesure généralement en visualisant les erreurs $\alpha(s)$ et $\beta(s)$ et/ou la spécificité et la sensibilité pour tous les seuils s.

Courbe ROC

- **Idée** : représenter sur un graphe 2d les deux types d'erreur pour tous les seuils s.

Définition

C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = \alpha(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = 1 - \beta(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

Remarque

- La courbe ROC d'un score parfait passe par le point (0,1).
- La courbe ROC d'un score aléatoire correspond à la première bissectrice.

Courbe ROC

- **Idée** : représenter sur un graphe 2d les deux types d'erreur pour tous les seuils s.

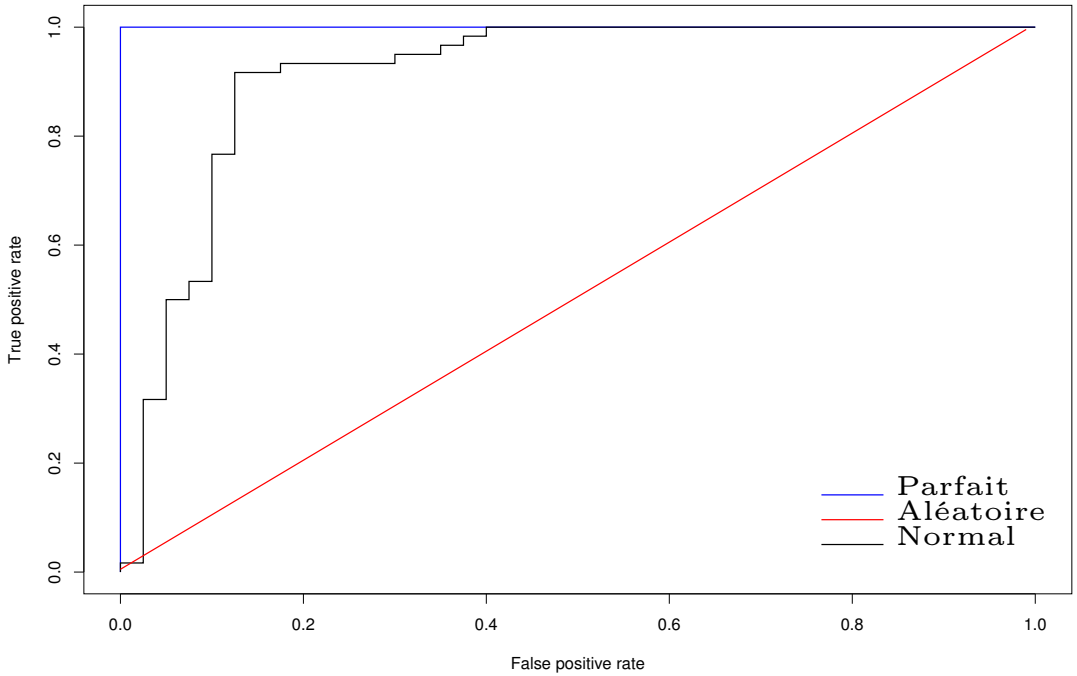
Définition

C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = \alpha(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = 1 - \beta(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

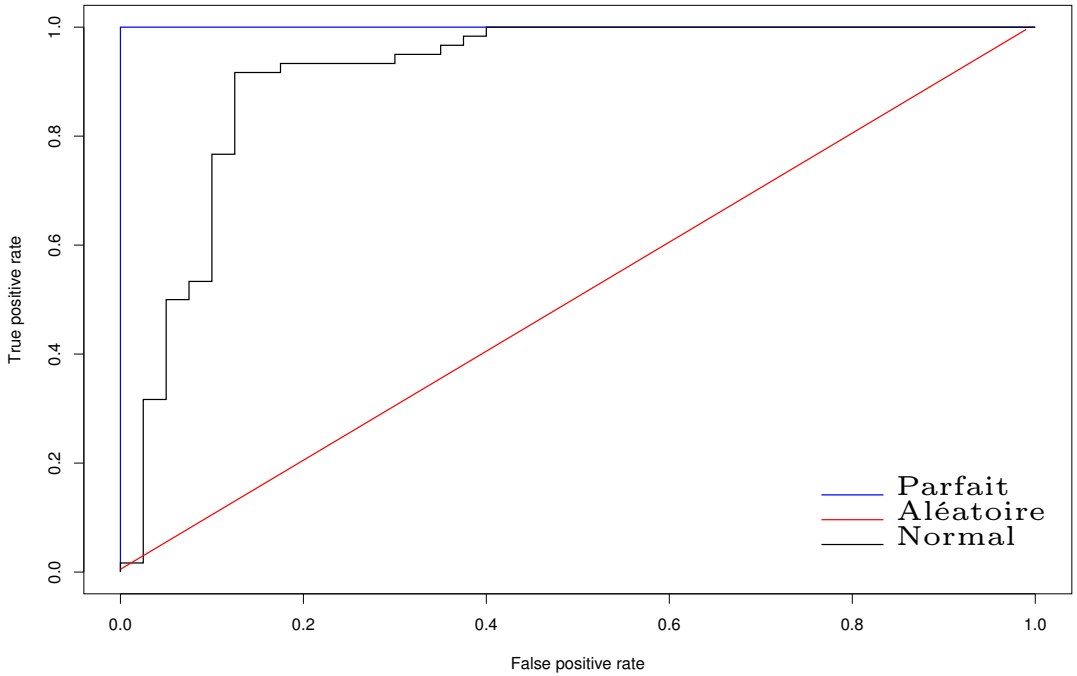
Remarque

- La courbe ROC d'un score parfait passe par le point (0,1).
- La courbe ROC d'un score aléatoire correspond à la première bissectrice.



Interprétation

On mesurera la performance d'un score par sa capacité à se rapprocher de la droite d'équation $y = 1$ le plus vite possible.



Interprétation

On mesurera la performance d'un score par sa capacité à se rapprocher de la droite d'équation $y = 1$ le plus vite possible.

AUC

Définition

- *L'aire sous la courbe ROC d'un score S, notée $AUC(S)$ est souvent utilisée pour mesurer sa performance.*
- *Pour un score parfait on a $AUC(S) = 1$, pour un score aléatoire $AUC(S) = 1/2$.*

Proposition

- *Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a*

$$AUC(S) = \mathbf{P}(S(X_1) \geq S(X_2) | (Y_1, Y_2) = (1, 0)).$$

AUC

Définition

- *L'aire sous la courbe ROC d'un score S, notée $AUC(S)$ est souvent utilisée pour mesurer sa performance.*
- *Pour un score parfait on a $AUC(S) = 1$, pour un score aléatoire $AUC(S) = 1/2$.*

Proposition

- *Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a*

$$AUC(S) = \mathbf{P}(S(X_1) \geq S(X_2) | (Y_1, Y_2) = (1, 0)).$$

Score optimal

- Le critère $AUC(S)$ peut être interprété comme une fonction de perte pour un score S ;
- Se pose donc la question d'existence d'un score optimal $S^\star$ vis-à-vis de ce critère.

Théorème ([Clémençon et al., 2008])

Soit $S^*(x) = \mathbf{P}(Y = 1|X = x)$, on a alors pour toutes fonctions de score S

$$AUC(S^*) \geq AUC(S).$$

Conséquence

Le problème pratique consistera à trouver un "bon" estimateur $S_n(x) = S_n(x, \mathcal{D}_n)$ de

$$S^*(x) = \mathbf{P}(Y = 1|X = x).$$

Score optimal

- Le critère $AUC(S)$ peut être interprété comme une fonction de perte pour un score S ;
- Se pose donc la question d'existence d'un score optimal $S^\star$ vis-à-vis de ce critère.

Théorème ([Clémençon et al., 2008])

Soit $S^\star(x) = \mathbf{P}(Y = 1|X = x)$, on a alors pour toutes fonctions de score S

$$AUC(S^\star) \geq AUC(S).$$

Conséquence

Le problème pratique consistera à trouver un "bon" estimateur $S_n(x) = S_n(x, \mathcal{D}_n)$ de

$$S^\star(x) = \mathbf{P}(Y = 1|X = x).$$

Score optimal

- Le critère $AUC(S)$ peut être interprété comme une fonction de perte pour un score S ;
- Se pose donc la question d'existence d'un score optimal $S^\star$ vis-à-vis de ce critère.

Théorème ([Clémençon et al., 2008])

Soit $S^\star(x) = \mathbf{P}(Y = 1|X = x)$, *on a alors pour toutes fonctions de score* S

$$AUC(S^\star) \geq AUC(S).$$

Conséquence

Le problème pratique consistera à trouver un "bon" estimateur $S_n(x) = S_n(x, \mathcal{D}_n)$ de

$$S^\star(x) = \mathbf{P}(Y = 1|X = x).$$

Score logistique

- de loin le plus utilisé...
- On considère le modèle logistique

$$\log \frac{p_\beta(x)}{1-p_\beta(x)} = \beta_1 x_1 + \ldots + \beta_p x_p$$

- Il suffit de poser $S(x) = p_\beta(x)$ ou

$$S(x) = \beta_1 x_1 + \ldots + \beta_p x_p.$$

Score logistique

- de loin le plus utilisé...
- On considère le modèle logistique

$$\log \frac{p_\beta(x)}{1-p_\beta(x)} = \beta_1 x_1 + \ldots + \beta_p x_p$$

- Il suffit de poser $S(x) = p_\beta(x)$ ou

$$S(x) = \beta_1 x_1 + \ldots + \beta_p x_p.$$

Illustration

- On calcule un score logistique sur l'exemple suivant :
  - On dispose d'un échantillon de taille $n = 150$ pour construire les fonctions de score (table dapp) :

	X1	X2	Y
1	-1.2070657	0.3158544	1
2	0.2774292	-2.1866448	0
3	1.0844412	-0.3307386	0
4	-2.3456977	-1.9001806	1
5	0.4291247	-0.3691092	0

  - On souhaite calculer le score pour 100 nouveaux individus (table dtest) :

	X1	X2
151	-0.37723765	-0.01545427
152	0.09761946	1.65997581
153	1.63874465	1.24334905
154	-0.87559247	-0.00564424
155	0.12176000	0.44504449

Illustration

- On calcule un score logistique sur l'exemple suivant :
  - On dispose d'un échantillon de taille $n = 150$ pour construire les fonctions de score (table `dapp`) :

```
           X1         X2 Y
1  -1.2070657  0.3158544 1
2   0.2774292 -2.1866448 0
3   1.0844412 -0.3307386 0
4  -2.3456977 -1.9001806 1
5   0.4291247 -0.3691092 0
```

  - On souhaite calculer le score pour 100 nouveaux individus (table `dtest`) :

```
             X1          X2
151 -0.37723765 -0.01545427
152  0.09761946  1.65997581
153  1.63874465  1.24334905
154 -0.87559247 -0.00564424
155  0.12176000  0.44504449
```

- On ajuste le modèle logistique sur l'échantillon d'apprentissage :

```
> model_logit <- glm(Y~.,data=dapp,family=binomial)
```

- On calcule le score des nouveaux individus :

```
> S1 <- predict(model_logit,newdata=dtest,type="response")
```

- On peut afficher le score de ces nouveaux individus :

```
> S1[1:5]
       151        152        153        154        155
0.77724343 0.56927363 0.02486394 0.92479413 0.51310825
```

- On ajuste le modèle logistique sur l'échantillon d'apprentissage :

```
> model_logit <- glm(Y~.,data=dapp,family=binomial)
```

- On calcule le score des nouveaux individus :

```
> S1 <- predict(model_logit,newdata=dtest,type="response")
```

- On peut afficher le score de ces nouveaux individus :

```
> S1[1:5]
       151        152        153        154        155
0.77724343 0.56927363 0.02486394 0.92479413 0.51310825
```

Autres modèles de score

- La plupart des modèles permettant d'expliquer une variables binaire par d'autres variables peuvent être utilisés pour construire des fonctions de score.
- Scores par arbre :

$$\hat{S}_T(x) = \hat{\mathsf{P}}(Y = 1 | X = x) = \frac{1}{n} \sum_{i: X_i \in \mathcal{N}(x)} \mathbf{1}_{Y_i = 1}.$$

- Scores LDA :

$$\hat{S}_{\text{LDA}}(x) = \hat{\delta}_1(x) = x^t \widehat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^t \widehat{\Sigma}^{-1} \hat{\mu}_1 + \log \hat{\pi}_1.$$

- Mais aussi SVM, boosting, forêts aléatoires...

Autres modèles de score

- La plupart des modèles permettant d'expliquer une variables binaire par d'autres variables peuvent être utilisés pour construire des fonctions de score.
- Scores par arbre :

$$\hat{S}_T(x) = \hat{\mathbf{P}}(Y = 1 | X = x) = \frac{1}{n} \sum_{i: X_i \in \mathcal{N}(x)} \mathbf{1}_{Y_i = 1}.$$

- Scores LDA :

$$\hat{S}_{\text{LDA}}(x) = \hat{\delta}_1(x) = x^t \widehat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^t \widehat{\Sigma}^{-1} \hat{\mu}_1 + \log \hat{\pi}_1.$$

- Mais aussi SVM, boosting, forêts aléatoires...

Autres modèles de score

- La plupart des modèles permettant d'expliquer une variables binaire par d'autres variables peuvent être utilisés pour construire des fonctions de score.
- Scores par arbre :

$$\hat{S}_T(x) = \hat{\mathsf{P}}(Y = 1 | X = x) = \frac{1}{n} \sum_{i: X_i \in \mathcal{N}(x)} \mathbf{1}_{Y_i = 1}.$$

- Scores LDA :

$$\hat{S}_{\text{LDA}}(x) = \hat{\delta}_1(x) = x^t \widehat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^t \widehat{\Sigma}^{-1} \hat{\mu}_1 + \log \hat{\pi}_1.$$

- Mais aussi SVM, boosting, forêts aléatoires...

Autres modèles de score

- La plupart des modèles permettant d'expliquer une variables binaire par d'autres variables peuvent être utilisés pour construire des fonctions de score.
- Scores par arbre :

$$\hat{S}_T(x) = \hat{\mathbf{P}}(Y = 1 | X = x) = \frac{1}{n} \sum_{i: X_i \in \mathcal{N}(x)} \mathbf{1}_{Y_i = 1}.$$

- Scores LDA :

$$\hat{S}_{\text{LDA}}(x) = \hat{\delta}_1(x) = x^t \widehat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^t \widehat{\Sigma}^{-1} \hat{\mu}_1 + \log \hat{\pi}_1.$$

- Mais aussi SVM, boosting, forêts aléatoires...

Comparaison de scores

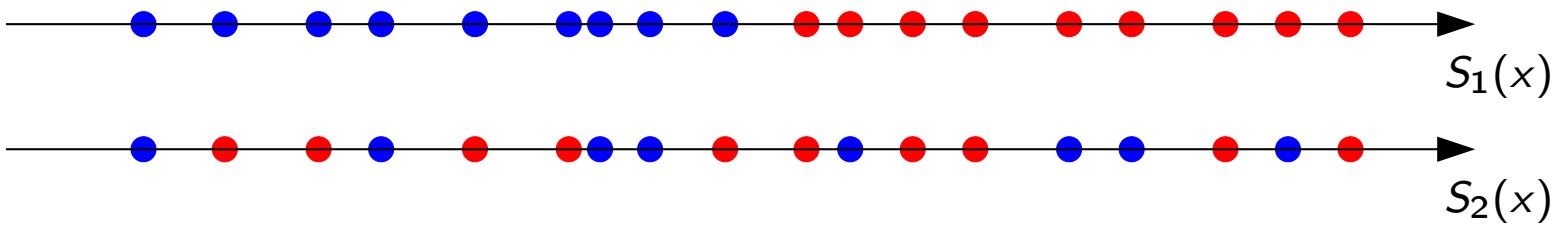
- On se trouve en présence de K fonctions de score $\hat{S}_1, \ldots, \hat{S}_K$.
- La comparaison des scores s'effectue généralement avec un échantillon (X_i, Y_i) indépendant de celui utilisé pour construire les fonctions de score $\hat{S}_k$.

Comparaison de scores

- On se trouve en présence de K fonctions de score $\hat{S}_1, \ldots, \hat{S}_K$.
- La comparaison des scores s'effectue généralement avec un échantillon (X_i, Y_i) indépendant de celui utilisé pour construire les fonctions de score $\hat{S}_k$.

Comparaison de scores

- On se trouve en présence de K fonctions de score $\hat{S}_1, \ldots, \hat{S}_K$.
- La comparaison des scores s'effectue généralement avec un échantillon (X_i, Y_i) indépendant de celui utilisé pour construire les fonctions de score $\hat{S}_k$.
- Graphiquement, on peut déjà avoir une première idée.



Courbe ROC (rappel)

- **Idée** : représenter sur un graphe 2d les deux types d'erreur pour tous les seuils s.

Définition

C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = \alpha(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = 1 - \beta(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

- On étudie l'allure de la courbe ROC à travers 2 scores particuliers :
  - Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que

$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad \text{et} \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

  - Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.

Courbe ROC (rappel)

- **Idée** : représenter sur un graphe 2d les deux types d'erreur pour tous les seuils s.

Définition

C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = \alpha(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = 1 - \beta(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

- On étudie l'allure de la courbe ROC à travers 2 scores particuliers :
  - Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que

$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad \text{et} \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

  - Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.

Courbe ROC (rappel)

- **Idée** : représenter sur un graphe 2d les deux types d'erreur pour tous les seuils s.

Définition

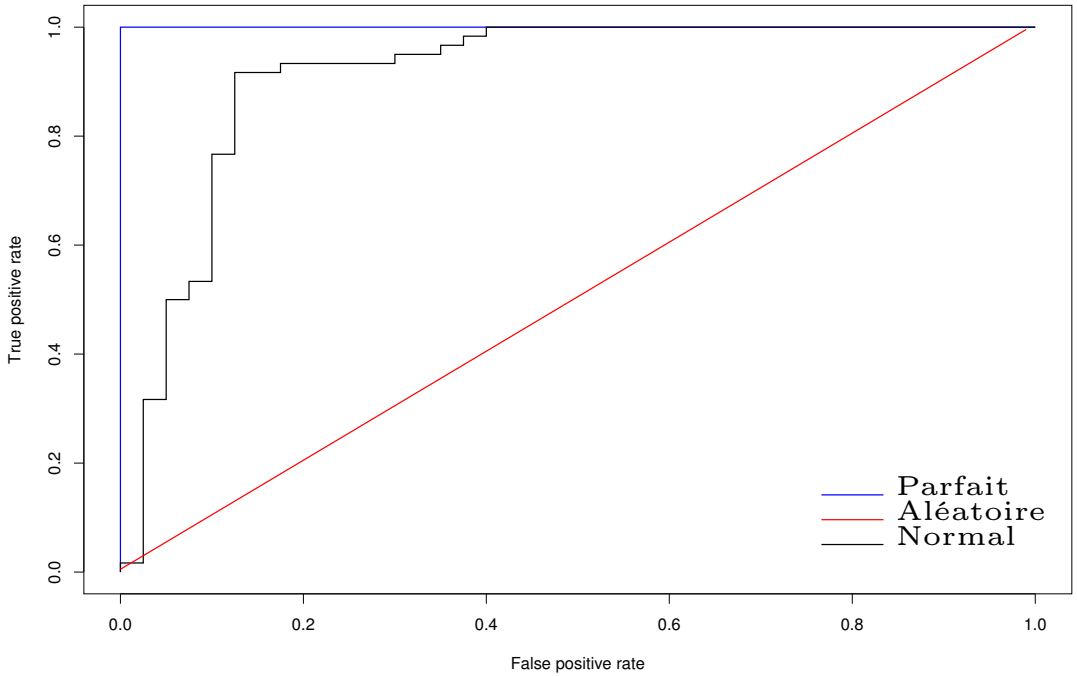
C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = \alpha(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = 1 - \beta(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

- On étudie l'allure de la courbe ROC à travers 2 scores particuliers :
  - Score parfait : il est tel qu'il existe un seuil $s^\star$ tel que

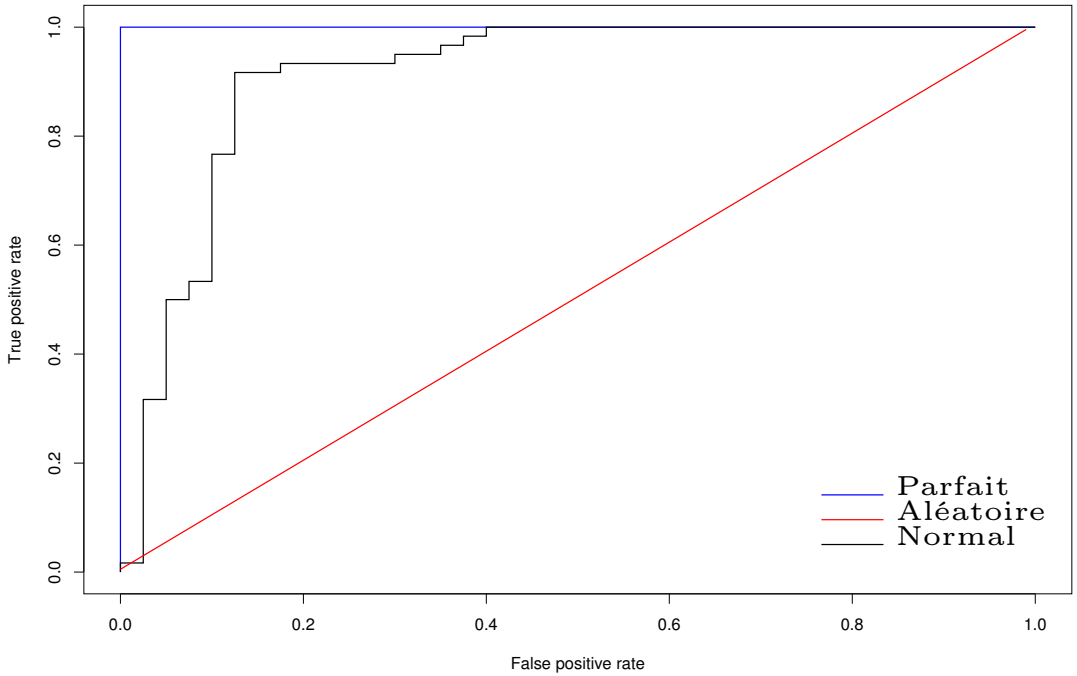
$$\mathbf{P}(Y = 1 | S(X) \geq s^\star) = 1 \quad \text{et} \quad \mathbf{P}(Y = 0 | S(X) < s^\star) = 1.$$

  - Score aléatoire : il est tel que $S(X)$ et Y sont indépendantes.



Interprétation

On mesurera la performance d'un score par sa capacité à se rapprocher de la droite d'équation $y = 1$ le plus vite possible.



Interprétation

On mesurera la performance d'un score par sa capacité à se rapprocher de la droite d'équation $y = 1$ le plus vite possible.

Courbe Lift

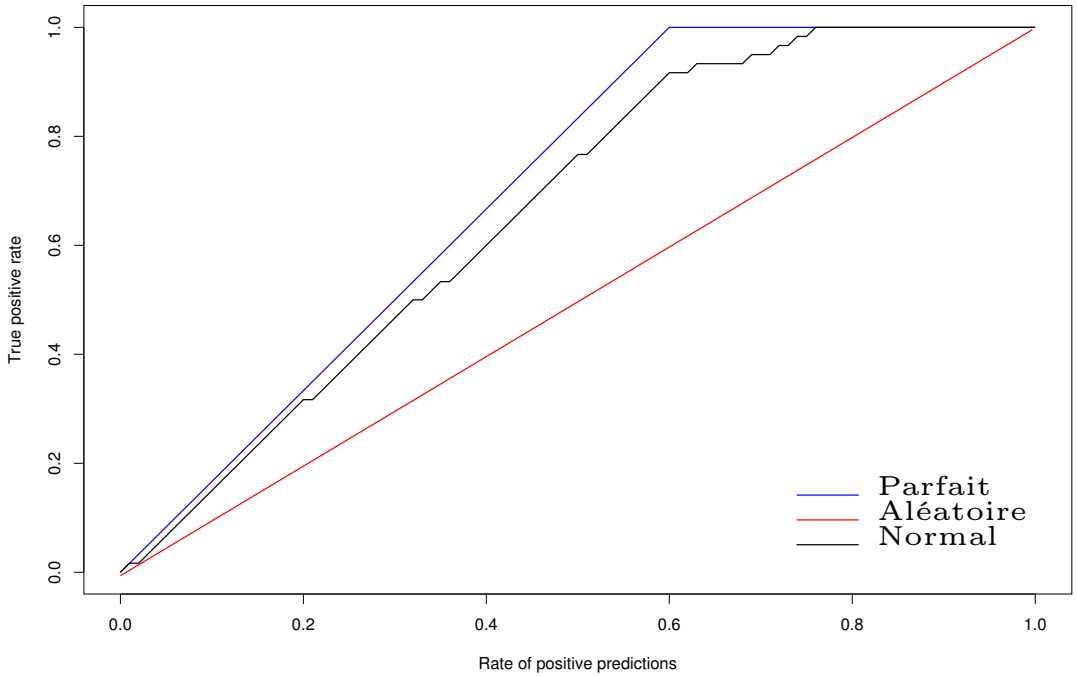
Définition

$$\begin{cases} x(s) = \mathbf{P}(S(X) \geq s) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

Courbe Lift

Définition

$$\begin{cases} x(s) = \mathbf{P}(S(X) \geq s) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$



Si je cible $x(s)$% des clients, je détecterai $y(s)$% des positifs.

Estimation

- Les courbes ROC et Lift nécessitent l'estimation de probabilités inconnues.
- **Exemple** : la courbe ROC associée à une fonction de score S nécessite le calcul des probabilités $\mathbf{P}(S(X) > s | Y = 0)$ et $\mathbf{P}(S(X) \geq s | Y = 1)$.
- L'estimation de ces quantités s'effectue à l'aide d'un échantillon indépendant de celui utilisé pour construire la fonction de score.

Estimation

- Les courbes ROC et Lift nécessitent l'estimation de probabilités inconnues.
- **Exemple** : la courbe ROC associée à une fonction de score S nécessite le calcul des probabilités $\mathbf{P}(S(X) > s | Y = 0)$ et $\mathbf{P}(S(X) \geq s | Y = 1)$.
- L'estimation de ces quantités s'effectue à l'aide d'un échantillon indépendant de celui utilisé pour construire la fonction de score.

Estimation

- Les courbes ROC et Lift nécessitent l'estimation de probabilités inconnues.
- **Exemple** : la courbe ROC associée à une fonction de score S nécessite le calcul des probabilités $\mathbf{P}(S(X) > s | Y = 0)$ et $\mathbf{P}(S(X) \geq s | Y = 1)$.
- L'estimation de ces quantités s'effectue à l'aide d'un échantillon indépendant de celui utilisé pour construire la fonction de score.

Apprentissage validation

- 2 quantités sont à estimer :
  1. La fonction de score $S(x)$
  2. Les paramètres de la courbe ROC : $\mathbf{P}(S(X) > s | Y = 0)$ et $\mathbf{P}(S(X) \geq s | Y = 1)$.

- L'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ est séparé deux :
  1. un échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$ utilisé pour estimer la fonction de score (par exemple les paramètres du modèle logistique pour le score logistique).
  2. un échantillon test $(X_{\ell+1}, Y_{\ell+1}), \ldots, (X_n, Y_n)$ pour estimer la courbe ROC

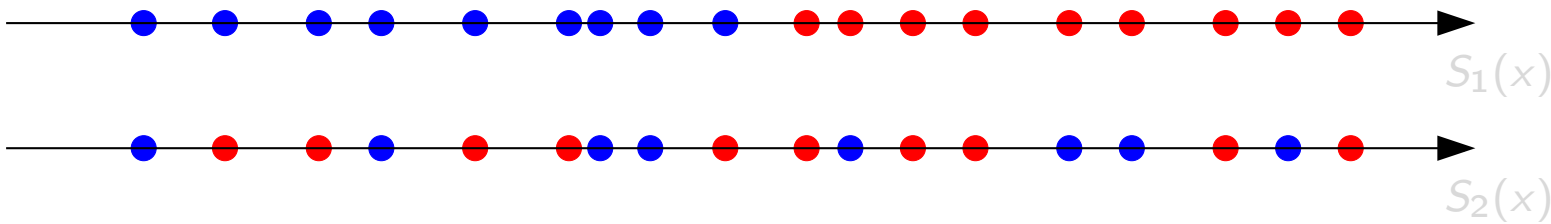
Apprentissage validation

- 2 quantités sont à estimer :
  1. La fonction de score $S(x)$
  2. Les paramètres de la courbe ROC : $\mathbf{P}(S(X) > s | Y = 0)$ et $\mathbf{P}(S(X) \geq s | Y = 1)$.

- L'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ est séparé deux :
  1. un échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$ utilisé pour estimer la fonction de score (par exemple les paramètres du modèle logistique pour le score logistique).
  2. un échantillon test $(X_{\ell+1}, Y_{\ell+1}), \ldots, (X_n, Y_n)$ pour estimer la courbe ROC

Une fois le score S estimé à l'aide de l'échantillon d'apprentissage, les paramètres de la courbe ROC sont estimés comme suit :

1. On applique le score aux variables explicatives de l'échantillon test.
2. On définit ainsi un nouvel échantillon $(S(X_{\ell+1}), Y_1), \ldots, (S(X_n), Y_n)$:



3. Les paramètres de la courbes ROC

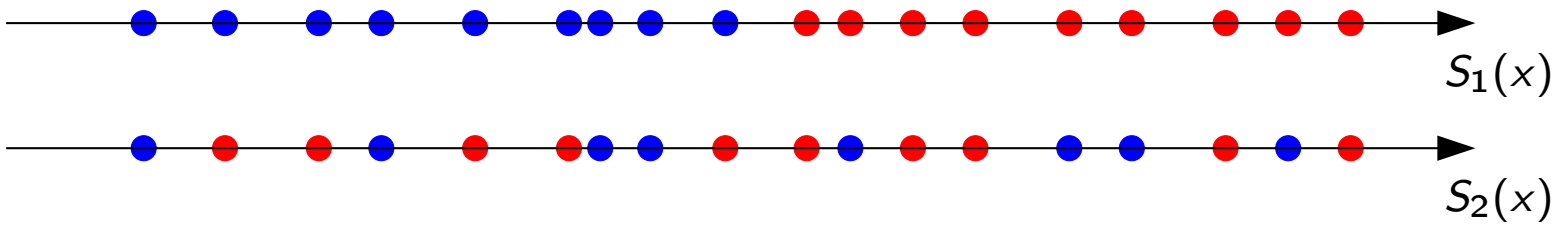
$$\begin{cases} x(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

sont estimés par

$$\begin{cases} \hat{x}(s) = \dfrac{1}{Card\{i : Y_i = 0\}} \displaystyle\sum_{i:Y_i=0} \mathbf{1}_{S(X_i)>s} \\ \\ \hat{y}(s) = \dfrac{1}{Card\{i : Y_i = 1\}} \displaystyle\sum_{i:Y_i=1} \mathbf{1}_{S(X_i)>s} \end{cases}$$

Une fois le score S estimé à l'aide de l'échantillon d'apprentissage, les paramètres de la courbe ROC sont estimés comme suit :

1. On applique le score aux variables explicatives de l'échantillon test.
2. On définit ainsi un nouvel échantillon $(S(X_{\ell+1}), Y_1), \ldots, (S(X_n), Y_n)$:



3. Les paramètres de la courbes ROC

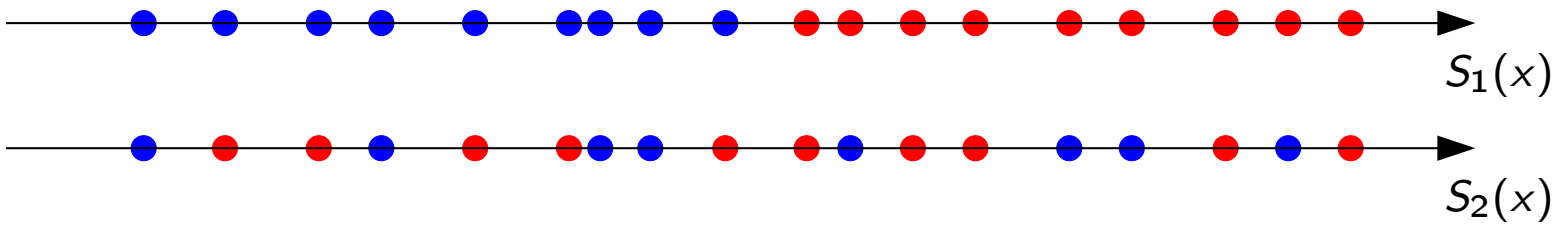
$$\begin{cases} x(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

sont estimés par

$$\begin{cases} \hat{x}(s) = \dfrac{1}{Card\{i : Y_i = 0\}} \displaystyle\sum_{i:Y_i=0} \mathbf{1}_{S(X_i)>s} \\ \\ \hat{y}(s) = \dfrac{1}{Card\{i : Y_i = 1\}} \displaystyle\sum_{i:Y_i=1} \mathbf{1}_{S(X_i)>s} \end{cases}$$

Une fois le score S estimé à l'aide de l'échantillon d'apprentissage, les paramètres de la courbe ROC sont estimés comme suit :

1. On applique le score aux variables explicatives de l'échantillon test.
2. On définit ainsi un nouvel échantillon $(S(X_{\ell+1}), Y_1), \ldots, (S(X_n), Y_n)$:



3. Les paramètres de la courbes ROC

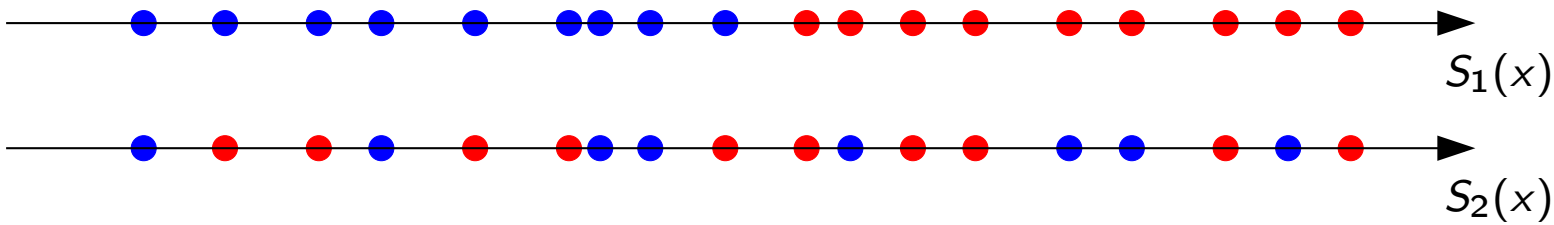
$$\begin{cases} x(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

sont estimés par

$$\begin{cases} \hat{x}(s) = \dfrac{1}{Card\{i : Y_i = 0\}} \displaystyle\sum_{i:Y_i=0} 1_{S(X_i)>s} \\ \hat{y}(s) = \dfrac{1}{Card\{i : Y_i = 1\}} \displaystyle\sum_{i:Y_i=1} 1_{S(X_i)>s} \end{cases}$$

Une fois le score S estimé à l'aide de l'échantillon d'apprentissage, les paramètres de la courbe ROC sont estimés comme suit :

1. On applique le score aux variables explicatives de l'échantillon test.
2. On définit ainsi un nouvel échantillon $(S(X_{\ell+1}), Y_1), \ldots, (S(X_n), Y_n)$:



3. Les paramètres de la courbes ROC

$$\begin{cases} x(s) = 1 - sp(s) = \mathbf{P}(S(X) > s | Y = 0) \\ y(s) = se(s) = \mathbf{P}(S(X) \geq s | Y = 1) \end{cases}$$

sont estimés par

$$\begin{cases} \hat{x}(s) = \dfrac{1}{Card\{i : Y_i = 0\}} \displaystyle\sum_{i:Y_i=0} \mathbf{1}_{S(X_i)>s} \\ \\ \hat{y}(s) = \dfrac{1}{Card\{i : Y_i = 1\}} \displaystyle\sum_{i:Y_i=1} \mathbf{1}_{S(X_i)>s} \end{cases}$$

Exemple

- On reprend l'exemple sur la maladie cardiovasculaire et on compare les 3 modèles construits (2 logistique et un arbre) à l'aide de la courbe ROC.
- On calcule d'abord le score des individus de l'échantillon test.

```
> score1 <- predict(model1,newdata=dtest,type="response")
> score2 <- predict(model2,newdata=dtest,type="response")
> score3 <- predict(arbre,newdata=dtest)
```

- On trace ensuite la courbe roc à l'aide de la fonction roc du package pROC.

```
> roc(dtest$chd,score1,plot=TRUE)
> roc(dtest$chd,score2,plot=TRUE,col="red",add=TRUE)
> roc(dtest$chd,score3,plot=TRUE,col="blue",add=TRUE)
> legend("bottomright",legend=c("logit1","logit2","arbre"),
                       col=c("black","red","blue"),lty=1,lwd=2)
```

Exemple

- On reprend l'exemple sur la maladie cardiovasculaire et on compare les 3 modèles construits (2 logistique et un arbre) à l'aide de la courbe ROC.
- On calcule d'abord le score des individus de l'échantillon test.

```
> score1 <- predict(model1,newdata=dtest,type="response")
> score2 <- predict(model2,newdata=dtest,type="response")
> score3 <- predict(arbre,newdata=dtest)
```

- On trace ensuite la courbe roc à l'aide de la fonction roc du package pROC.

```
> roc(dtest$chd,score1,plot=TRUE)
> roc(dtest$chd,score2,plot=TRUE,col="red",add=TRUE)
> roc(dtest$chd,score3,plot=TRUE,col="blue",add=TRUE)
> legend("bottomright",legend=c("logit1","logit2","arbre"),
                       col=c("black","red","blue"),lty=1,lwd=2)
```

Exemple

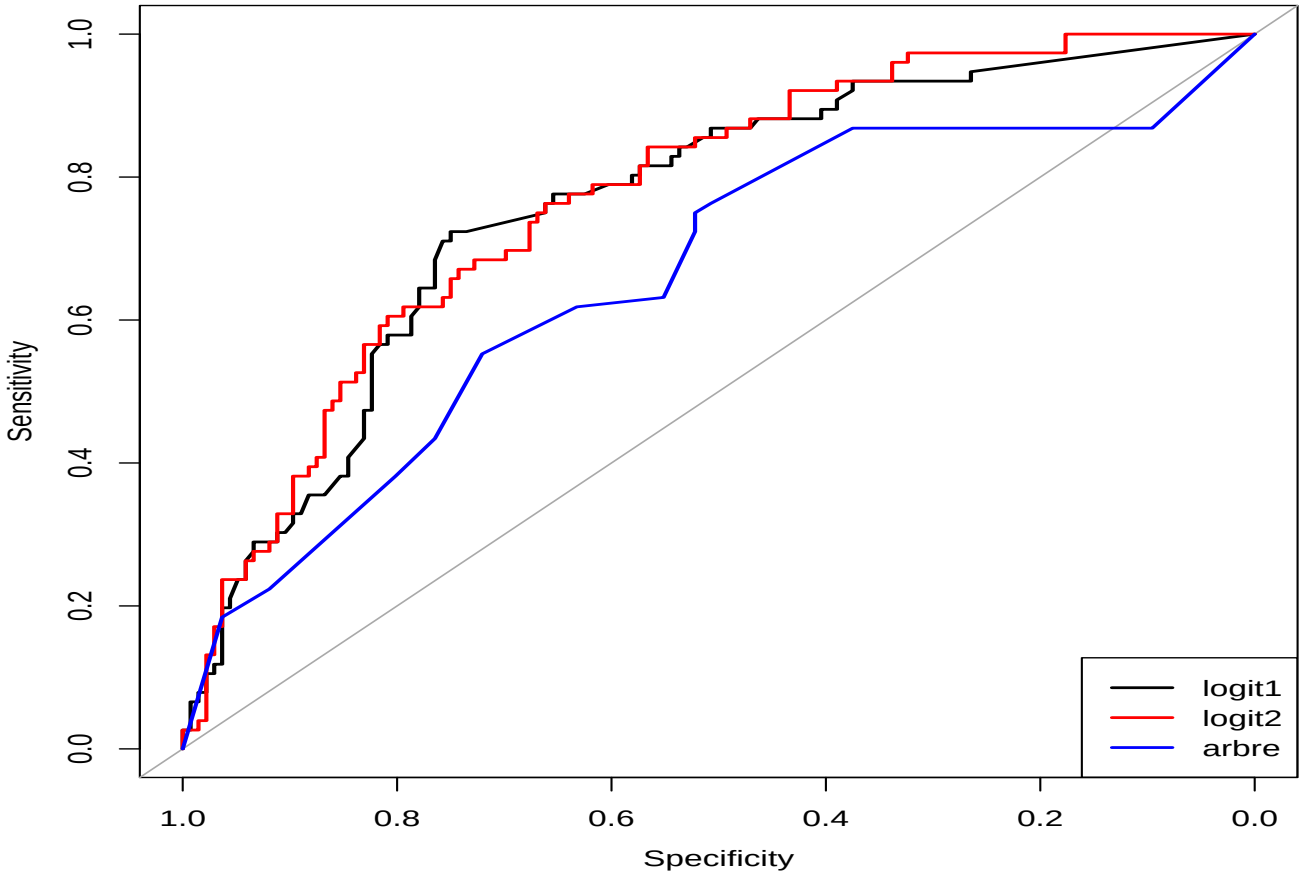
- On reprend l'exemple sur la maladie cardiovasculaire et on compare les 3 modèles construits (2 logistique et un arbre) à l'aide de la courbe ROC.
- On calcule d'abord le score des individus de l'échantillon test.

```
> score1 <- predict(model1,newdata=dtest,type="response")
> score2 <- predict(model2,newdata=dtest,type="response")
> score3 <- predict(arbre,newdata=dtest)
```

- On trace ensuite la courbe roc à l'aide de la fonction roc du package pROC.

```
> roc(dtest$chd,score1,plot=TRUE)
> roc(dtest$chd,score2,plot=TRUE,col="red",add=TRUE)
> roc(dtest$chd,score3,plot=TRUE,col="blue",add=TRUE)
> legend("bottomright",legend=c("logit1","logit2","arbre"),
                       col=c("black","red","blue"),lty=1,lwd=2)
```

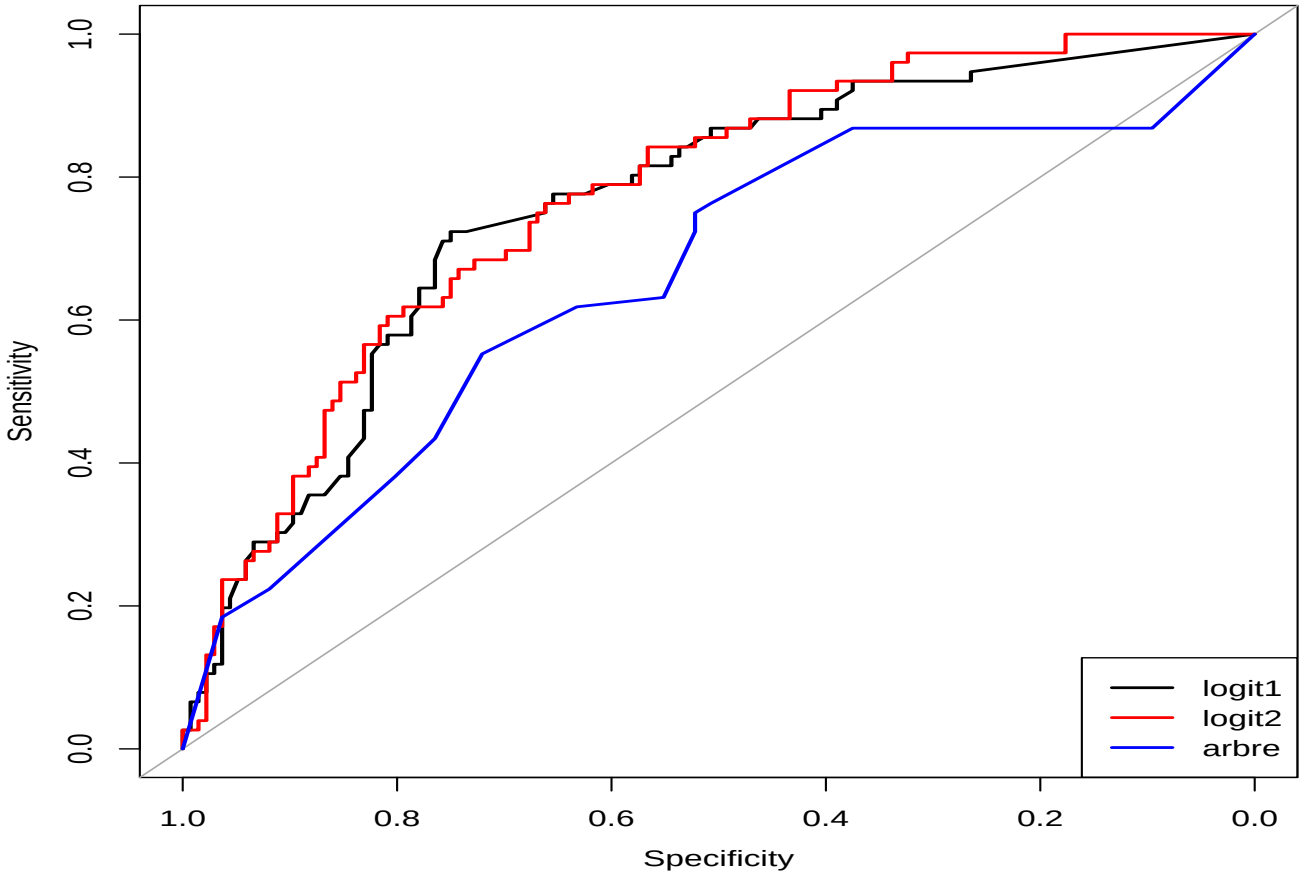
Courbes ROC



Conclusion

Pour le critère ROC, les modèles logistique sont plus performants que l'arbre de classification.

Courbes ROC



Conclusion

Pour le critère ROC, les modèles logistique sont plus performants que l'arbre de classification.

- La plupart sont basés sur les indicateurs graphique.

AUC et AUL

- Ils représentent respectivement l'aire sous la courbe ROC et l'aire sous la courbe lift.
- On a

$$AUC - AUL = p(AUC - 0.5) \iff AUL = p/2 + (1-p)AUC$$

avec $p = \mathbf{P}(Y = 1)$.

- Score parfait : $AUC = 1$, $AUL = 1 - p/2$.
- Score aléatoire : $AUC = AUL = 1/2$.

Propriété

Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a

$$AUC = \mathbf{P}(S(X_1) > S(X_2) | (Y_1, Y_2) = (1, 0)).$$

- La plupart sont basés sur les indicateurs graphique.

AUC et AUL

- Ils représentent respectivement l'aire sous la courbe ROC et l'aire sous la courbe lift.
- On a

$$AUC - AUL = p(AUC - 0.5) \iff AUL = p/2 + (1-p)AUC$$

avec $p = \mathbf{P}(Y = 1)$.

- Score parfait : $AUC = 1$, $AUL = 1 - p/2$.
- Score aléatoire : $AUC = AUL = 1/2$.

Propriété

Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a

$$AUC = \mathbf{P}(S(X_1) > S(X_2) | (Y_1, Y_2) = (1, 0)).$$

- La plupart sont basés sur les indicateurs graphique.

AUC et AUL

- Ils représentent respectivement l'aire sous la courbe ROC et l'aire sous la courbe lift.
- On a

$$AUC - AUL = p(AUC - 0.5) \iff AUL = p/2 + (1-p)AUC$$

avec $p = \mathbf{P}(Y = 1)$.

- Score parfait : $AUC = 1$, $AUL = 1 - p/2$.
- Score aléatoire : $AUC = AUL = 1/2$.

Propriété

Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a

$$AUC = \mathbf{P}(S(X_1) > S(X_2) | (Y_1, Y_2) = (1, 0)).$$

- La plupart sont basés sur les indicateurs graphique.

AUC et AUL

- Ils représentent respectivement l'aire sous la courbe ROC et l'aire sous la courbe lift.
- On a

$$AUC - AUL = p(AUC - 0.5) \iff AUL = p/2 + (1-p)AUC$$

avec $p = \mathbf{P}(Y = 1)$.

- Score parfait : $AUC = 1$, $AUL = 1 - p/2$.
- Score aléatoire : $AUC = AUL = 1/2$.

Propriété

Etant données deux observations (X_1, Y_1) et (X_2, Y_2) indépendantes et de même loi que (X, Y), on a

$$AUC = \mathbf{P}(S(X_1) > S(X_2) | (Y_1, Y_2) = (1, 0)).$$

Tau de Kendall

- Il permet de mesurer la relation entre deux scores S_1 et S_2 en se basant sur la manière dont les deux scores classent n individus.

Définition

Une paire (X_i, X_j) est concordante pour deux scores S_1 et S_2 si

$$S_1(X_i) < S_1(X_j) \quad \text{et} \quad S_2(X_i) < S_2(X_j)$$

ou

$$S_1(X_i) > S_1(X_j) \quad \text{et} \quad S_2(X_i) > S_2(X_j).$$

Sinon la paire est dite discordante.

Définition

Si $\forall i \neq j$, $S_1(X_i) \neq S_1(X_j)$ et $S_2(X_i) \neq S_2(X_j)$, alors le τ de Kendall est défini par

$$\tau_K = \frac{\text{nb de paires concor.} - \text{nb de paires discor.}}{n(n-1)/2}.$$

Propriété

- $-1 \leq \tau_K \leq 1$
- $\tau_K = 1 \iff S_1$ et S_2 font le même classement.
- $\tau_K = -1 \iff S_1$ et S_2 font le classement opposé.

- Le τ de Kendall peut aussi être utilisé pour mesurer la performance d'un score S.
- Dans ce cas, on confronte $S(X_1), \ldots, S(X_n)$ à $Y_1, \ldots, Y_n$.
- La formule précédente n'est pas applicable mais il existe des variantes permettant de conserver la même interprétation pour les valeurs de τ_K.

- Le τ de Kendall peut aussi être utilisé pour mesurer la performance d'un score S.
- Dans ce cas, on confronte $S(X_1), \ldots, S(X_n)$ à $Y_1, \ldots, Y_n$.
- La formule précédente n'est pas applicable mais il existe des variantes permettant de conserver la même interprétation pour les valeurs de τ_K.

Les données

- On souhaite ici comparer différents modèles de scores via le critère ROC pour les données spam.
- On scinde l'échantillon en
  1. un échantillon d'apprentissage de taille 2300 pour ajuster les différents modèles.
  2. un échantillon test de taille 2301 pour mesurer la performance des scores (estimer les courbes ROC).

```
> donnees <- read.csv("spam.csv")
> napp <- 2300
> indapp <- 1:napp
> dapp <- donnees[indapp,]
> dtest <- donnees[-indapp,]
```

Les données

- On souhaite ici comparer différents modèles de scores via le critère ROC pour les données spam.
- On scinde l'échantillon en
  1. un échantillon d'apprentissage de taille 2300 pour ajuster les différents modèles.
  2. un échantillon test de taille 2301 pour mesurer la performance des scores (estimer les courbes ROC).

```
> donnees <- read.csv("spam.csv")
> napp <- 2300
> indapp <- 1:napp
> dapp <- donnees[indapp,]
> dtest <- donnees[-indapp,]
```

Les données

- On souhaite ici comparer différents modèles de scores via le critère ROC pour les données spam.
- On scinde l'échantillon en
  1. un échantillon d'apprentissage de taille 2300 pour ajuster les différents modèles.
  2. un échantillon test de taille 2301 pour mesurer la performance des scores (estimer les courbes ROC).

```
> donnees <- read.csv("spam.csv")
> napp <- 2300
> indapp <- 1:napp
> dapp <- donnees[indapp,]
> dtest <- donnees[-indapp,]
```

Construction des modèles candidats

- On met en concurrence les modèles logistiques, lda, forêts aléatoires et arbres :

```
> logit <- glm(Y~.,data=dapp,family=binomial)
> lda1 <- lda(Y~.,data=dapp)
> RF <- randomForest(Y~.,data=dapp)
> arbre <- rpart(Y~.,data=dapp)
```

- Puis on calcule, pour chaque modèle, le score des individus de l'échantillon test :

```
> S_logit <- predict(logit,newdata=dtest)
> S_lda <- predict(lda1,newdata=dtest)$x
> S_RF <- predict(RF,newdata=dtest,type="prob")[,2]
> S_arbre <- predict(arbre,newdata=dtest,tpe="prob")[,2]
```

Construction des modèles candidats

- On met en concurrence les modèles logistiques, lda, forêts aléatoires et arbres :

```
> logit <- glm(Y~.,data=dapp,family=binomial)
> lda1 <- lda(Y~.,data=dapp)
> RF <- randomForest(Y~.,data=dapp)
> arbre <- rpart(Y~.,data=dapp)
```

- Puis on calcule, pour chaque modèle, le score des individus de l'échantillon test :

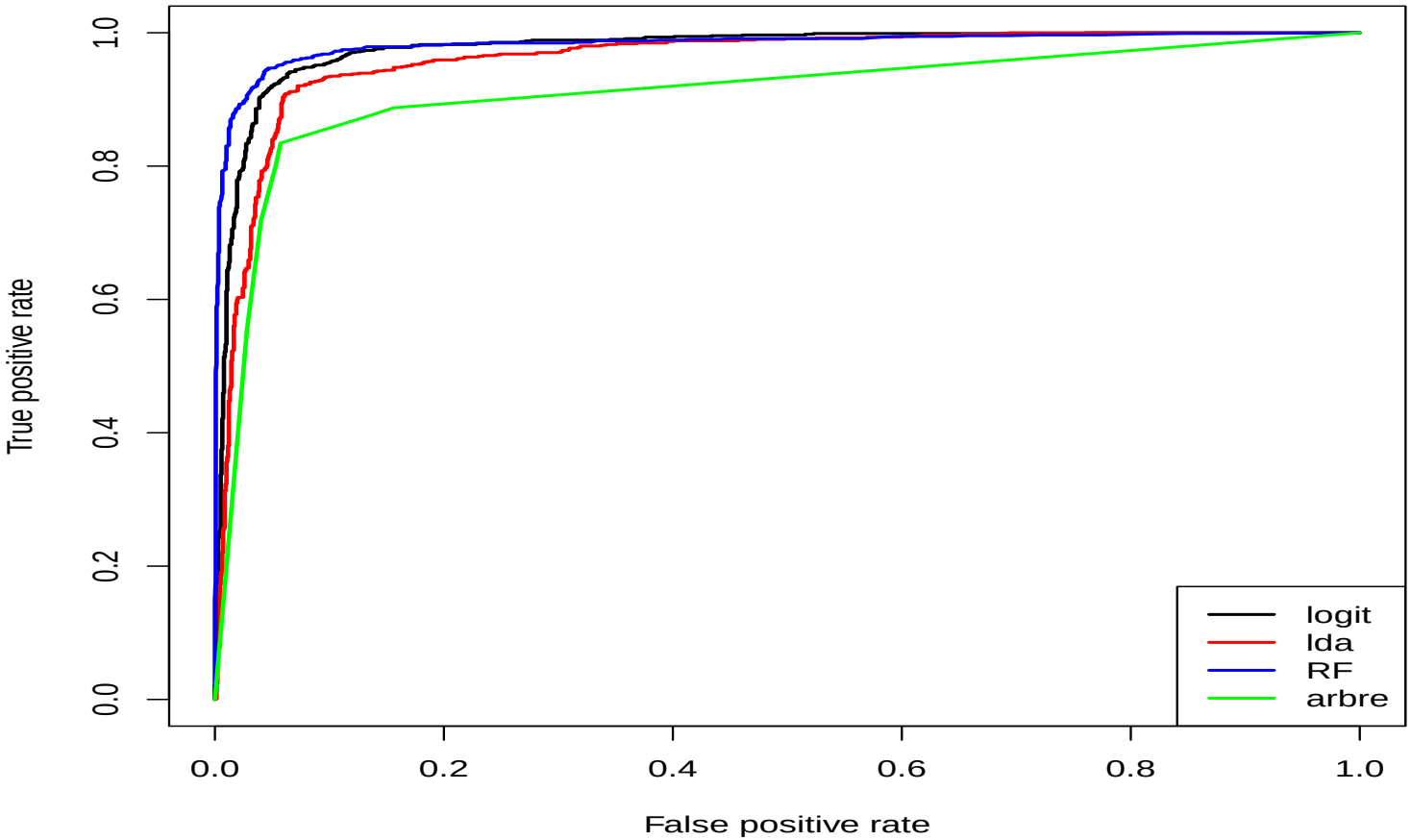
```
> S_logit <- predict(logit,newdata=dtest)
> S_lda <- predict(lda1,newdata=dtest)$x
> S_RF <- predict(RF,newdata=dtest,type="prob")[,2]
> S_arbre <- predict(arbre,newdata=dtest,tpe="prob")[,2]
```

Construction des courbes ROC

- On trace les 4 courbes ROC estimées.

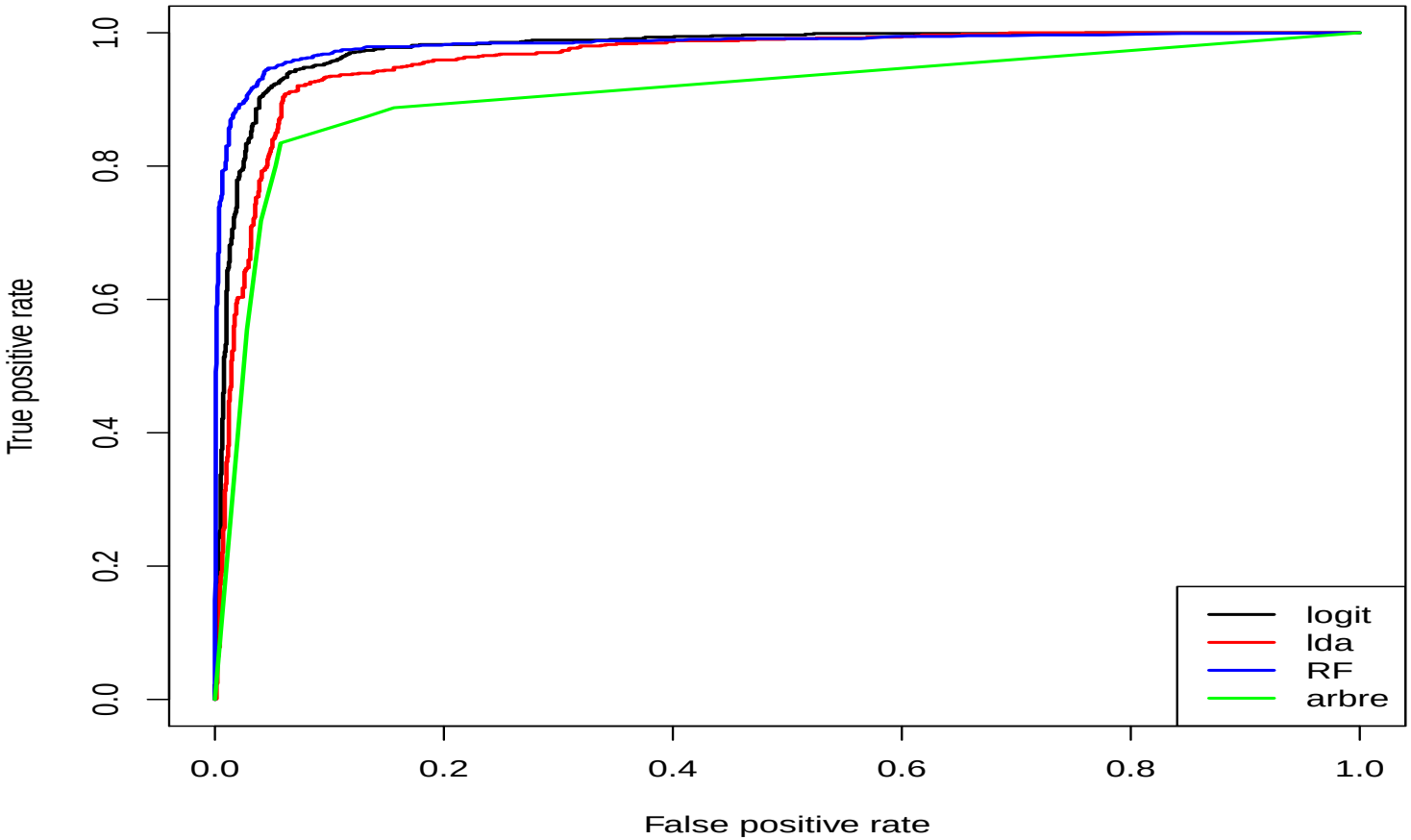
```
> library(ROCR)
> S1_pred <- prediction(S_logit,dtest$Y)
> S2_pred <- prediction(S_lda,dtest$Y)
> S3_pred <- prediction(S_RF,dtest$Y)
> S4_pred <- prediction(S_arbre,dtest$Y)
>
> roc1 <- performance(S1_pred,measure="tpr",x.measure="fpr")
> roc2 <- performance(S2_pred,measure="tpr",x.measure="fpr")
> roc3 <- performance(S3_pred,measure="tpr",x.measure="fpr")
> roc4 <- performance(S4_pred,measure="tpr",x.measure="fpr")
>
> plot(roc1,col="black",lwd=2)
> plot(roc2,add=TRUE,col="red",lwd=2)
> plot(roc3,add=TRUE,col="blue",lwd=2)
> plot(roc4,add=TRUE,col="green",lwd=2)
> legend("bottomright",legend=c("logit","lda","RF","arbre"),
            col=c("black","red","blue","green"),lty=1,lwd=2)
```

Courbes ROC



Pour le critère ROC, on privilégiera le score par forêt aléatoire.

Courbes ROC



Pour le critère ROC, on privilégiera le score par forêt aléatoire.

AUC et τ_K

- On obtient sur R l'AUC et le τ_K du score logistique avec

```
> performance(S1_pred,"auc")@y.values[[1]]
[1] 0.9772703
> Kendall(S_logit,dtest$Y)
tau = 0.66, 2-sided pvalue =< 2.22e-16
```

- On peut ainsi comparer les 4 scores

	Logit	LDA	Forêt	Arbre
AUC	0.977	0.961	0.983	0.910
τ_K	0.66	0.637	0.672	0.717

AUC et τ_K

- On obtient sur R l'AUC et le τ_K du score logistique avec

```
> performance(S1_pred,"auc")@y.values[[1]]
[1] 0.9772703
> Kendall(S_logit,dtest$Y)
tau = 0.66, 2-sided pvalue =< 2.22e-16
```

- On peut ainsi comparer les 4 scores

	Logit	LDA	Forêt	Arbre
AUC	0.977	0.961	0.983	0.910
τ_K	0.66	0.637	0.672	0.717

Références I


Clémençon, S., Lugosi, G., and Vayatis, N. (2008).
Ranking and empirical minimization of u-statistics.
The Annals of Statistics, 36(2) :844–874.