Data Science

M2 Actuariat

Session 2 : Modèles Linéaires Généralisés et Pénalisés

François HU

Responsable du pôle Intelligence Artificielle, Milliman R&D

Enseignant ISFA

2025

ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES

Milliman

Sommaire du cours Data Science

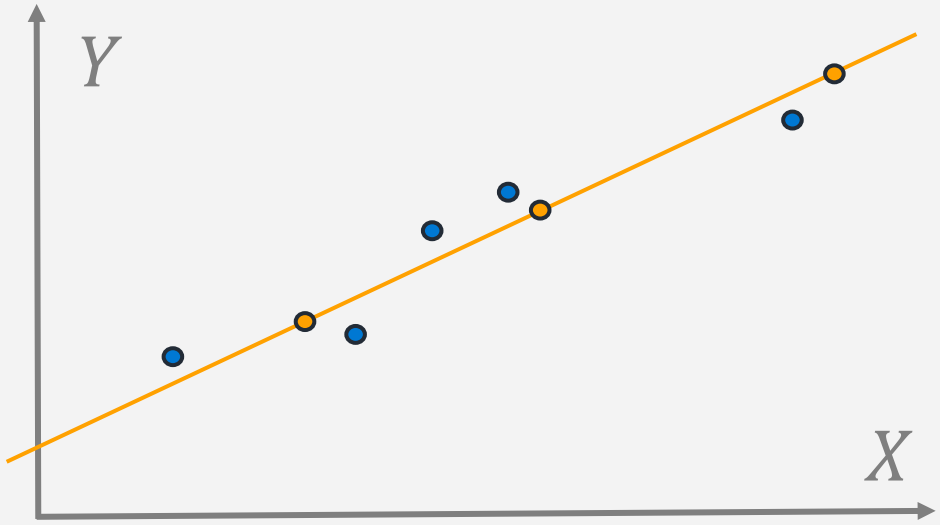
1. Apprentissage statistique et lien avec l'Actuariat
2. Modèles linéaires généralisés et pénalisés
3. Arbre de décision et méthodes ensemblistes
4. Interprétabilité des modèles d'apprentissage
5. IA de confiance et biais algorithmiques
6. Apprentissage non supervisé
7. Introduction aux données non structurées

Session modèles linéaires généralisés et pénalisés

1. Modèles Linéaires : Rappel et Limites
2. Composants des modèles GLM : fonctions de lien et distributions de probabilité
3. Modèle Logistique et Poisson
4. Estimation des paramètres : La méthode de vraisemblance
5. Interprétation des coefficients
6. Evaluation des modèles GLM
7. Sélection des variables explicatives
8. GLM pénalisé : Régression Lasso et Ridge
9. Extension des modèles GLM aux modèles GAM

1. Modèles Linéaires

- **Rappel**
- Limites



Milliman

Rappel des modèles linéaires simples (1/5)

Définitions et notations

$$Y = \beta_0 + \beta X + \varepsilon$$

Variable cible ou variable à expliquer, ou variable dépendante

Coefficient(s) ou paramètres

Variable(s) **explicative**(s) ou prédicteur(s)

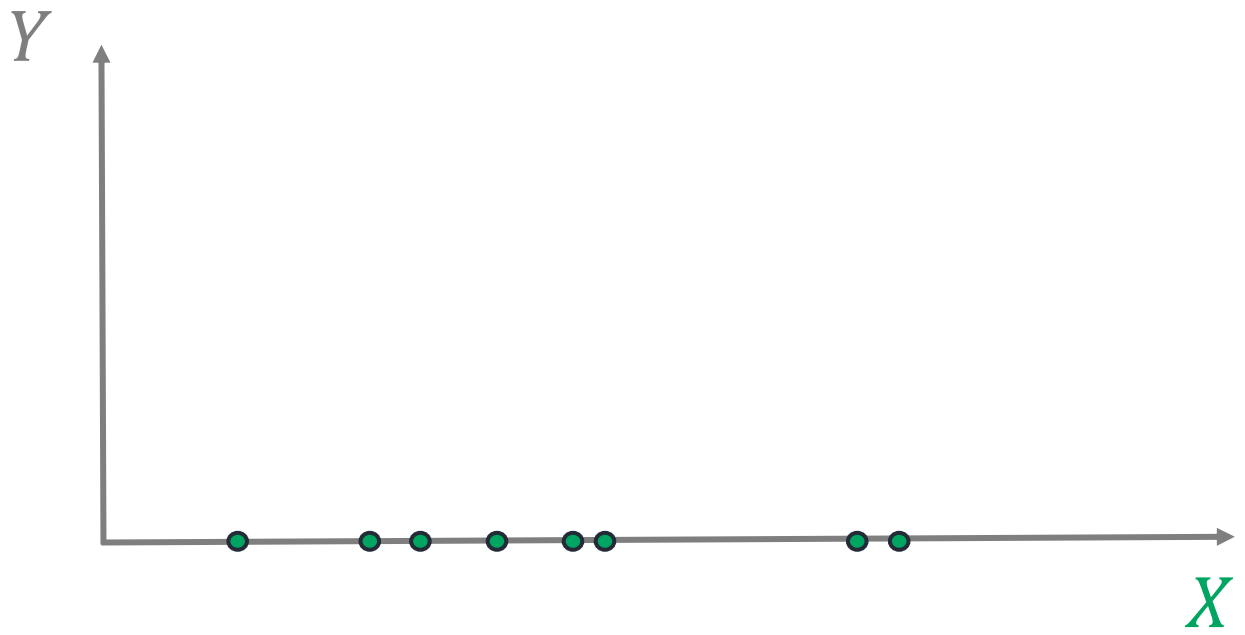
Résidu ou **erreur aléatoire**, supposée suivre une distribution normale avec une moyenne zéro

Les **modèles linéaires simples** permettent de décrire la relation entre deux variables quantitatives (**variable cible** Y et **variable explicative** X) par une équation linéaire.

Pour un **portefeuille automobile**, la variable cible peut être le coût du sinistre (en supposant que la personne est exposée) et la variable explicative l'âge du conducteur.

Le **principal objectif** du modèle linéaire est de **prédire** les valeurs Y à partir de X en estimant les **coefficients β et β_0**. Les coefficients sont généralement estimés par la **méthode des moindres carrées**.

Exemple illustratif



Rappel des modèles linéaires simples (2/5)

Evaluer la généralisation

$$Y = \beta_0 + \beta X + \varepsilon$$

Variable cible ou variable à expliquer, ou variable dépendante

Coefficient(s) ou paramètres

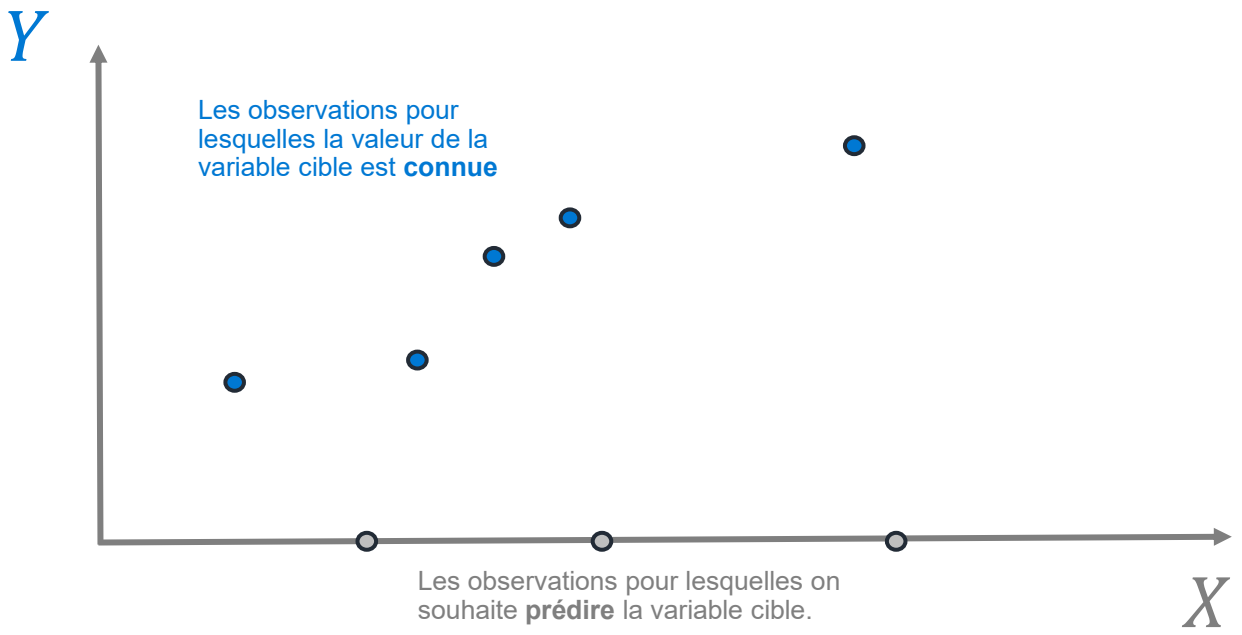
Variable(s) **explicative**(s) ou prédicteur(s)

Résidu ou **erreur aléatoire**, supposée suivre une distribution normale avec une moyenne zéro

Séparation train-test : La séparation entre les ensembles d'entraînement (train) et de test est importante pour évaluer la performance réelle d'un modèle linéaire

Évaluer la généralisation : Le modèle linéaire est ajusté sur l'ensemble d'entraînement (**train**), les données qu'il utilise pour apprendre. Cependant, cela ne garantit pas qu'il fonctionnera bien sur de nouvelles données, non vues auparavant. L'ensemble de **test**, composé de données que le modèle n'a jamais vues, permet de vérifier la capacité du modèle à **généraliser** à des données inconnues.

Exemple illustratif



Rappel des modèles linéaires simples (3/5)

Différence entre entraînement et inférence

$$Y = \beta_0 + \beta X + \varepsilon$$

Variable cible ou variable à expliquer, ou variable dépendante

Coefficient(s) ou paramètres

Variable explicative(s) ou prédicteur(s)

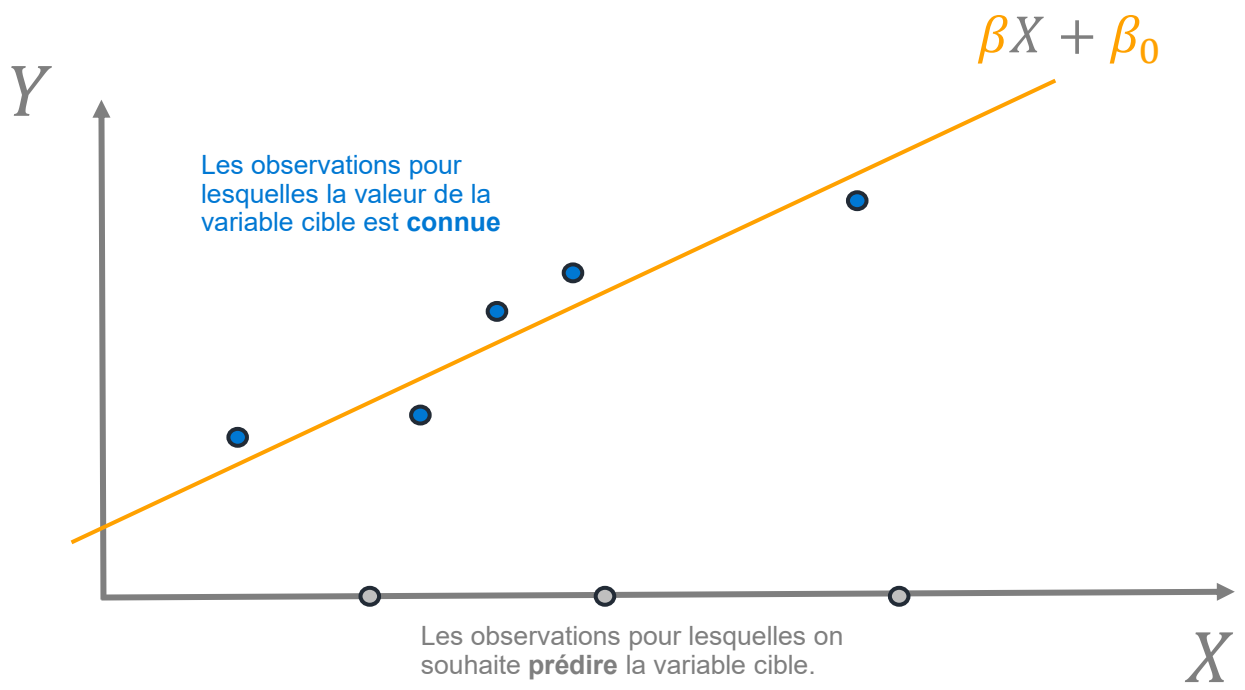
Résidu ou **erreur aléatoire**, supposée suivre une distribution normale avec une moyenne zéro

L'entraînement consiste à ajuster les **paramètres** d'un modèle à partir de **données connues** pour qu'il apprenne des patterns.

L'inférence, en revanche, est le processus où le modèle, une fois entraîné, fait des prédictions sur des nouvelles **données non vues**.

- En résumé, **l'entraînement** est l'apprentissage, tandis que **l'inférence** est l'utilisation du modèle pour prédire.

Exemple illustratif



Rappel des modèles linéaires simples (4/5)

Hypothèses d'un modèle linéaire

$$Y = \beta_0 + \beta X + \varepsilon$$

Variable cible ou variable à expliquer, ou variable dépendante

Coefficient(s) ou paramètres

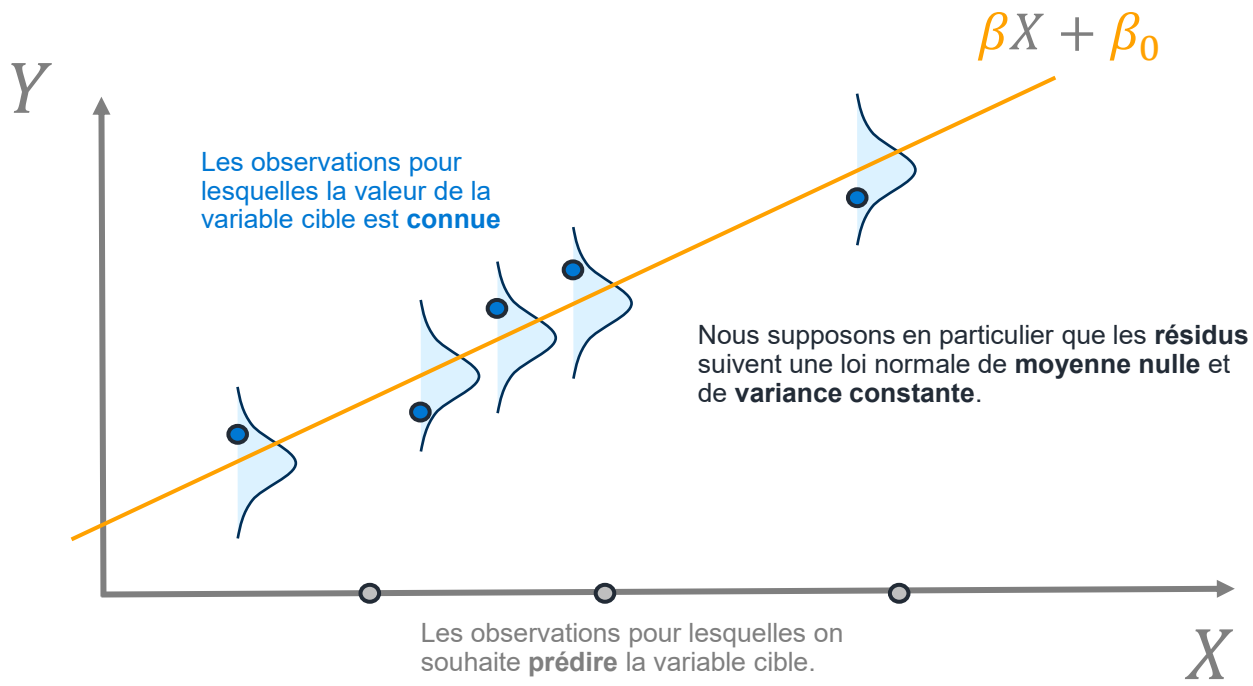
Variable explicative(s) ou prédicteur(s)

Résidu ou **erreur aléatoire**, supposée suivre une distribution normale avec une moyenne zéro

Les 4 hypothèses fondamentales **des erreurs ε_i** (qui doivent être vérifiées pour un modèle linéaire) :

1) **Moyenne nulle (erreurs centrées) :** le modèle ne doit pas présenter de biais systématique.
2) **Variance constante (homoscédasticité) :** La dispersion des erreurs doit être constante, pour éviter que certaines parties du modèle soient plus imprécises que d'autres.
3) **Indépendance des erreurs :** l'erreur commise pour une observation ne doit pas influencer les autres.
4) **Distribution normale (loi gaussienne) :** Les erreurs suivent une distribution normale, ce qui permet d'appliquer les tests statistiques et d'estimer la précision des prédictions.

Exemple illustratif



Rappel des modèles linéaires simples (5/5)

Performance du modèle

$$Y = \beta_0 + \beta X + \varepsilon$$

Variable cible ou variable à expliquer, ou variable dépendante

Coefficient(s) ou paramètres

Variable explicative(s) ou prédicteur(s)

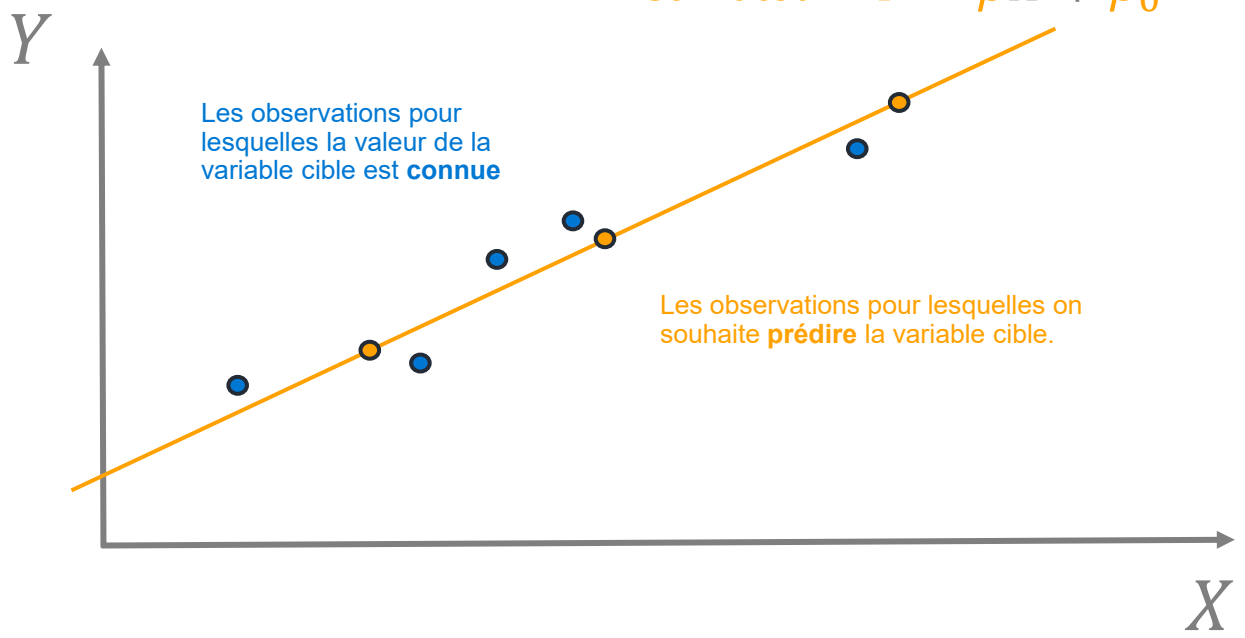
Résidu ou **erreur aléatoire**, supposée suivre une distribution normale avec une moyenne zéro

Plusieurs approches pour vérifier si un modèle linéaire est bien estimé :

- **Visuellement :** Utiliser des graphiques comme le QQ-plot pour vérifier si les erreurs suivent une distribution normale.
- **Empiriquement :** Utiliser le coefficient de détermination R^2 pour mesurer la proportion de la variation expliquée par le modèle par rapport à la variation totale.
- **Statistiquement :** Effectuer une analyse de la variance (ANOVA) pour évaluer la significativité du modèle avec la statistique F et sa p-valeur.

Exemple illustratif

Estimateur : $\hat{Y} = \beta X + \beta_0$



Modèles linéaires multiples (1/5)

Plusieurs variables explicatives : des modèles linéaires dites « multiples »

Exemple illustratif

- Cas avec **plusieurs variables explicatives** (souvent plus de 10 variables).

Coût sinistre	*Catégorie*	*Ancienneté*	*Age*
0	3	0	41
122.3	1	8	20
977.8	2	7	18
0	1	5	47
311.1	1	2	29

Modèles linéaires multiples (2/5)

Un peu de formalisme

- **Modèles linéaires multiples** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

$$Y = \begin{bmatrix} 0 \\ 122.3 \\ 977.8 \\ 0 \\ 311.1 \end{bmatrix} \quad X_1 = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 1 \\ 1 \end{bmatrix} \quad X_2 = \begin{bmatrix} 0 \\ 8 \\ 7 \\ 5 \\ 2 \end{bmatrix} \quad X_3 = \begin{bmatrix} 41 \\ 20 \\ 18 \\ 47 \\ 29 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

Modèles linéaires multiples (3/5)

Ecriture matricielle

Modèles linéaires multiples $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

- **Ecriture matricielle**

$$\underbrace{\begin{bmatrix} 0 \\ 122.3 \\ 977.8 \\ 0 \\ 311.1 \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & 3 & 0 & 41 \\ 1 & 1 & 8 & 20 \\ 1 & 2 & 7 & 18 \\ 1 & 1 & 5 & 47 \\ 1 & 1 & 2 & 29 \end{bmatrix}}_{\mathbf{X}} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\boldsymbol{\beta}^T} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

Modèles linéaires multiples (4/5)

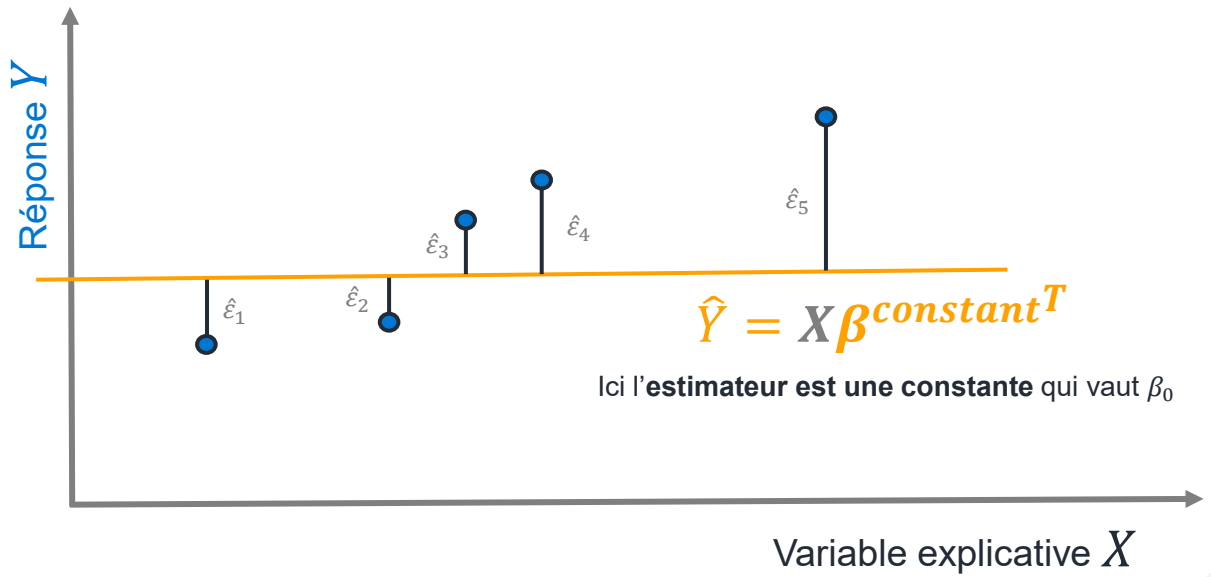
Estimateur constant

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

Notre objectif est d'estimer les coefficients $\boldsymbol{\beta}$. Plus précisément nous souhaitons trouver $\boldsymbol{\beta}$ qui minimise la méthode des moindres carrés, une approche statistique utilisée pour ajuster un modèle à des données en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle

$$R(\boldsymbol{\beta}) \coloneqq \sum_i \left(Y_i - X_i \beta_i^T\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

Exemple d'un mauvais estimateur avec $R(\boldsymbol{\beta}^{best})$ grand



Ici l'**estimateur est une constante** qui vaut β_0

Modèles linéaires multiples (5/5)

Estimateur optimal

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

Notre objectif est d'estimer les coefficients $\boldsymbol{\beta}$. Plus précisément nous souhaitons trouver $\boldsymbol{\beta}$ qui minimise la méthode des moindres carrés, une approche statistique utilisée pour ajuster un modèle à des données en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle

$$R(\boldsymbol{\beta}) := \sum_i \left(Y_i - X_i \beta_i^T\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

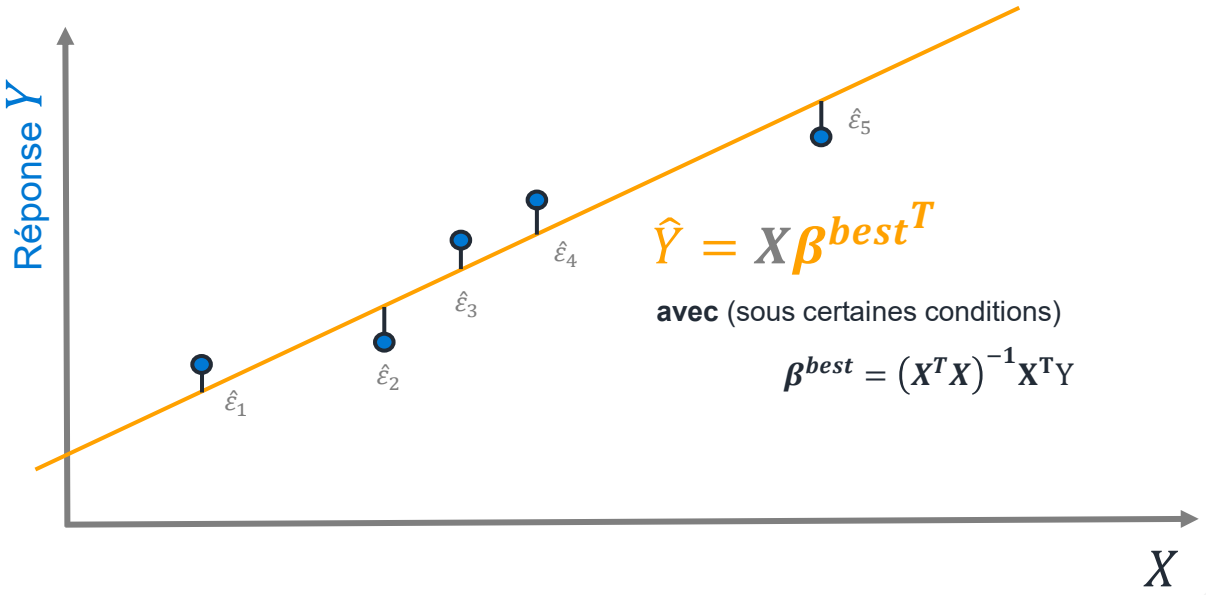
Expression explicite du paramètre optimal β^{best} : si toutes les variables explicatives dans $\boldsymbol{X}$ sont **linéairement indépendantes** entre elles alors

$$\boldsymbol{\beta^{best}} = \left(\boldsymbol{X^T X}\right)^{-1} \mathbf{X^T} \mathrm{Y}$$

avec $\mathbb{E}[\boldsymbol{\beta^{best}}] = \boldsymbol{\beta}$ (on dit que l'estimateur est non biaisé)

$\mathbb{V}[\boldsymbol{\beta^{best}}] = \sigma^2 (\boldsymbol{X^T X})^{-1}$ la matrice variance-covariance

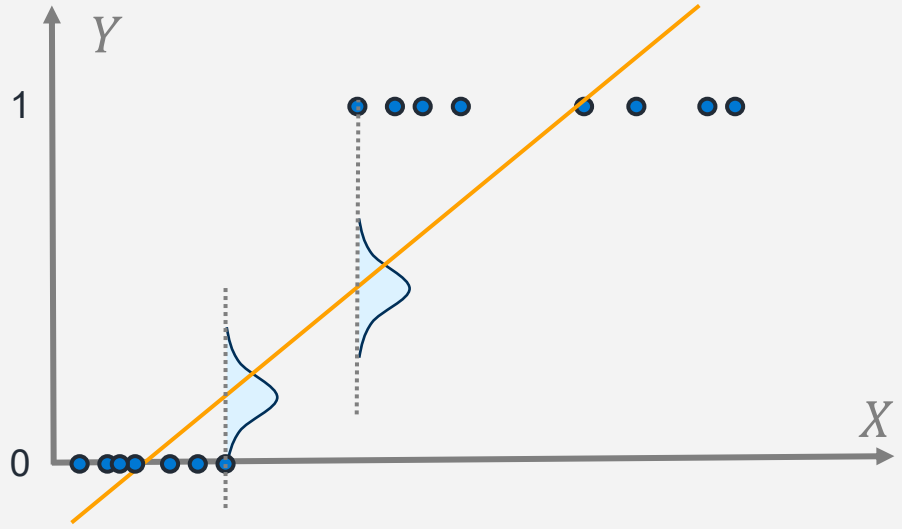
Exemple d'un bon estimateur avec $R(\boldsymbol{\beta^{best}})$ petit



1. Modèles Linéaires

Rappel

- **Limites**



ISFA
Institut de Science Financière et d'Assurances

Limites des modèles linéaires classiques

Cas variable cible **binaire**

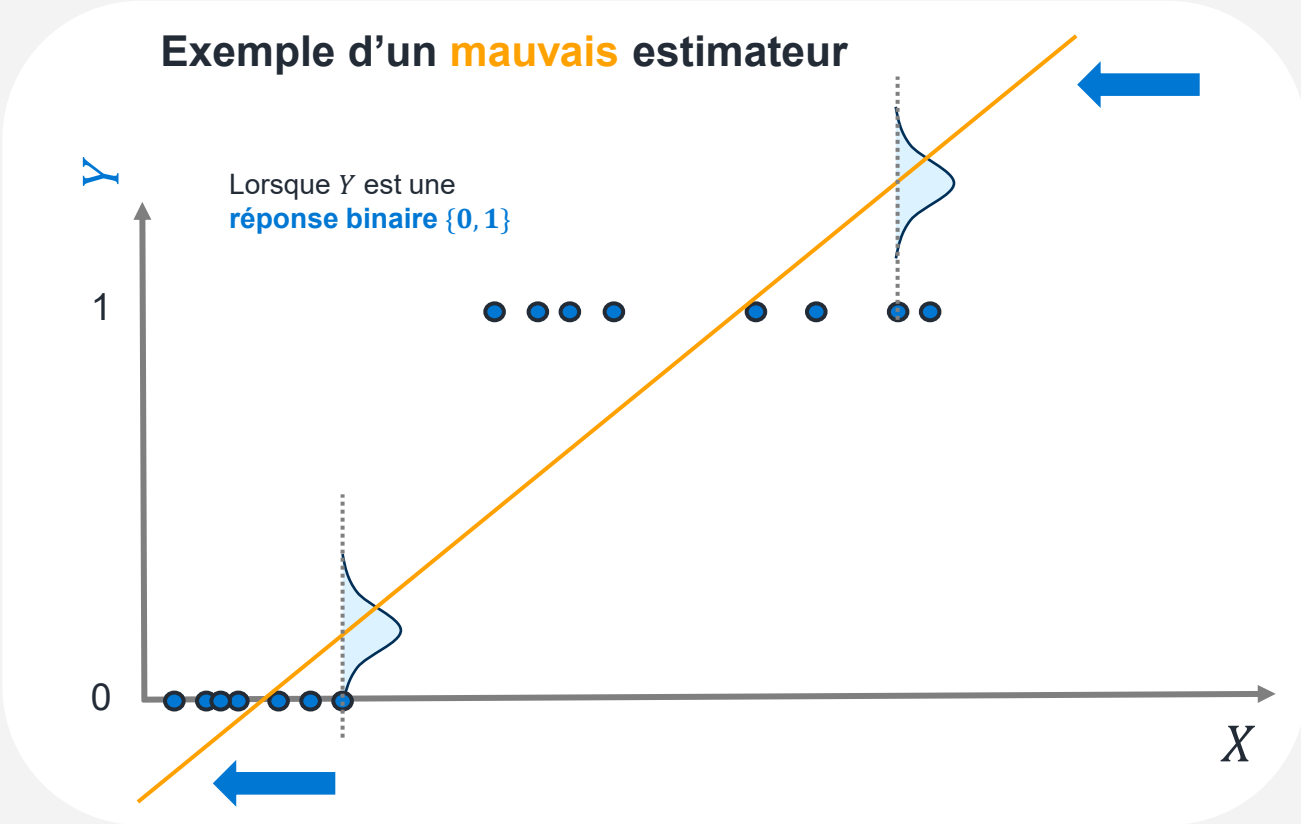
Exemple illustratif

▶ **Modèle linéaire multiple** $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

Lorsque Y n'est pas une variable cible continue alors le **modèle linéaire classique** n'a plus de sens :

1) L'estimateur à pour valeur dans un **ensemble discret**
2) Les **hypothèses sur les résidus** (notamment de variance constante) ne sont pas vérifiées



Limites des modèles linéaires classiques

Cas variable cible **binaire**

Exemple illustratif

▶ **Solution : Prédire la probabilité d'un événement binaire** (0 ou 1) à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

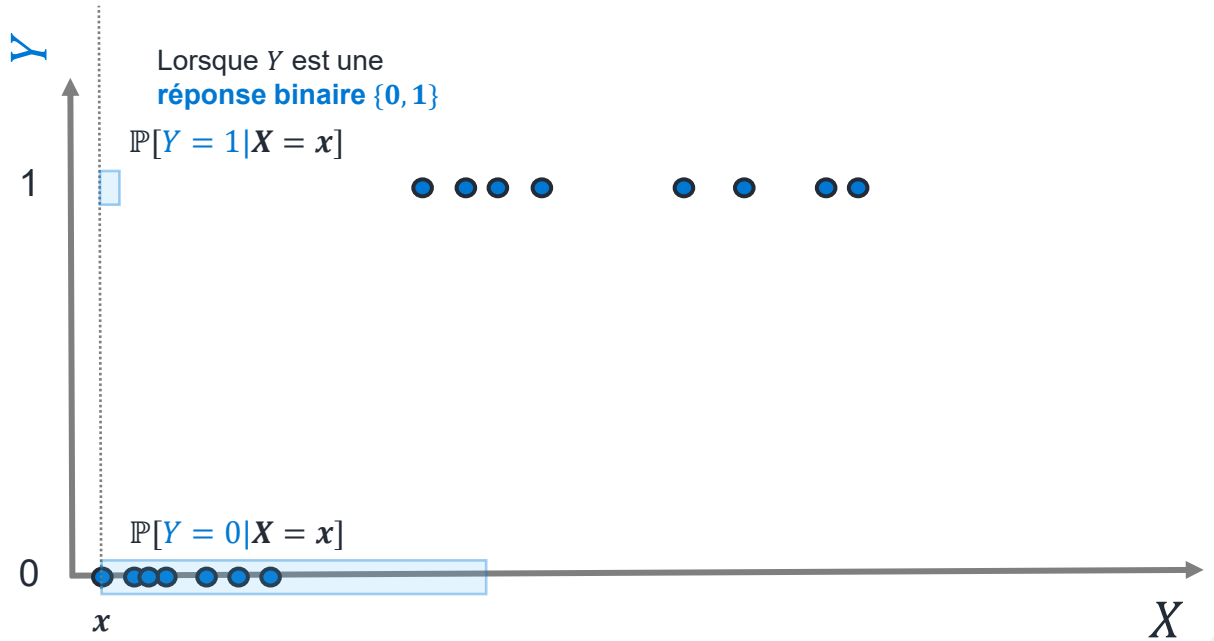
Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{T}}$~~

Réponse binaire : $\mathbb{E}[Y|\boldsymbol{X}] = \mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \boldsymbol{p}$

Distribution « naturelle » : $Y \mid \boldsymbol{X} \sim Ber(\boldsymbol{p})$

Lien ? $\boldsymbol{p} \leftrightarrow \boldsymbol{\beta}$

Objectif



Limites des modèles linéaires classiques

Cas variable cible **binaire**

Exemple illustratif

▶ **Solution : Prédire la probabilité d'un événement binaire** (0 ou 1) à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

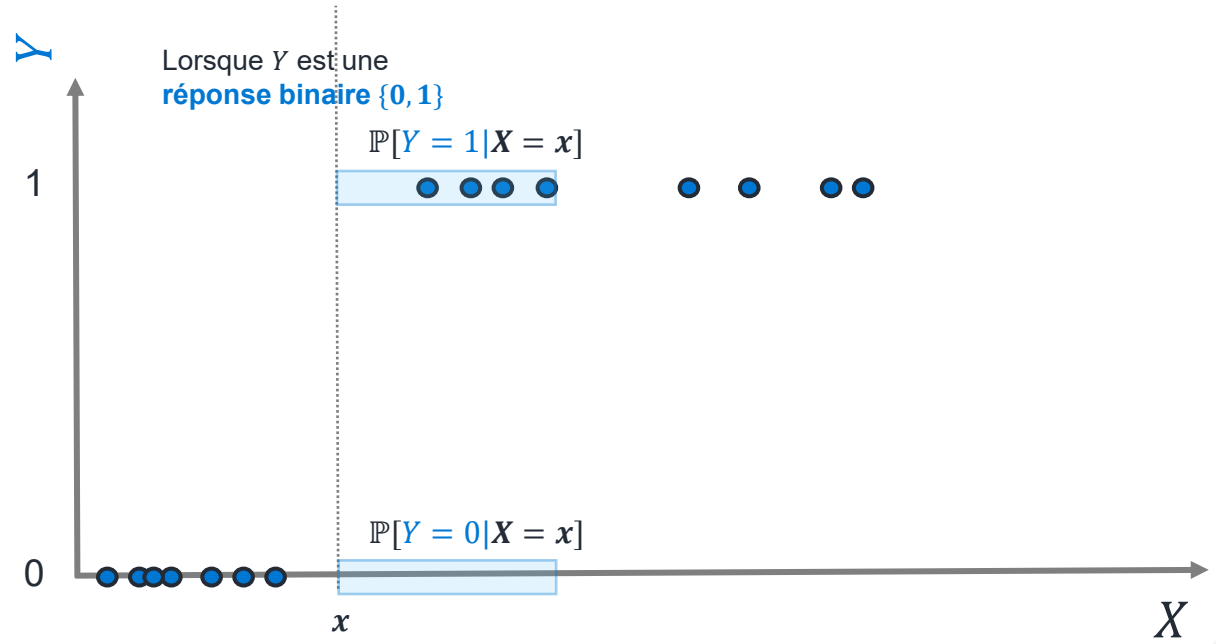
Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{T}}$~~

Réponse binaire : $\mathbb{E}[Y|\boldsymbol{X}] = \mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \boldsymbol{p}$

Distribution « naturelle » : $Y \mid \boldsymbol{X} \sim Ber(\boldsymbol{p})$

Lien ? $\boldsymbol{p} \leftrightarrow \boldsymbol{\beta}$

Objectif



Limites des modèles linéaires classiques

Cas variable cible **binaire**

Exemple illustratif

▶ **Solution : Prédire la probabilité d'un événement binaire** (0 ou 1) à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

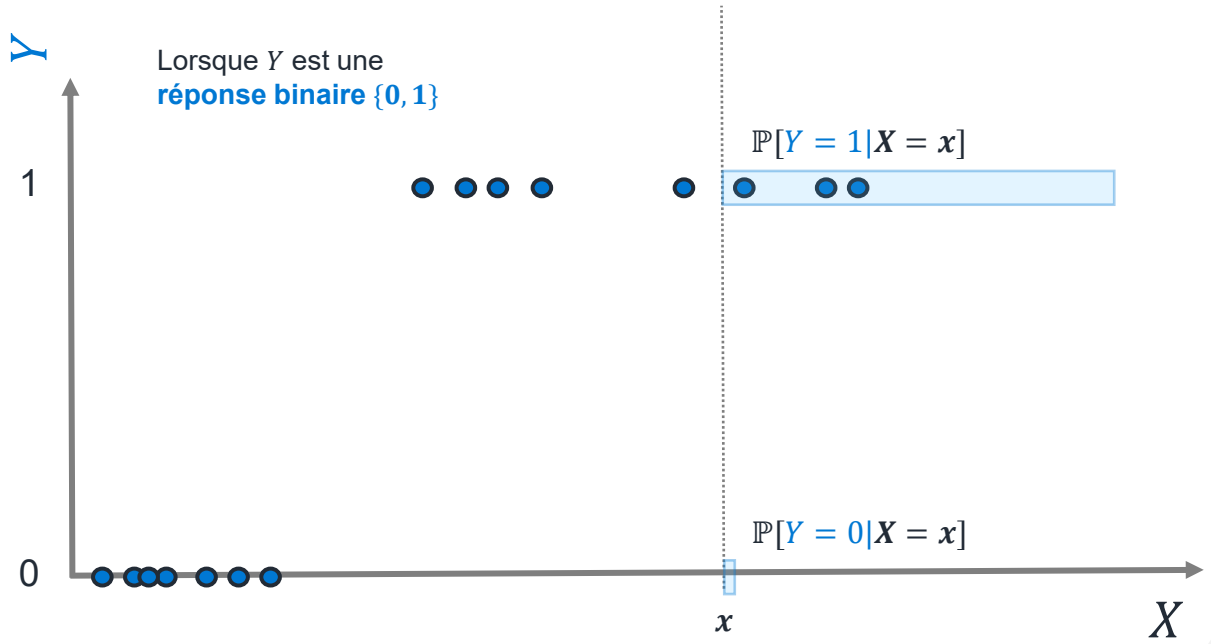
Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{T}}$~~

Réponse binaire : $\mathbb{E}[Y|\boldsymbol{X}] = \mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \boldsymbol{p}$

Distribution « naturelle » : $Y \mid \boldsymbol{X} \sim Ber(\boldsymbol{p})$

Lien ? $\boldsymbol{p} \leftrightarrow \boldsymbol{\beta}$

Objectif



Limites des modèles linéaires classiques

Cas variable cible **discrète**

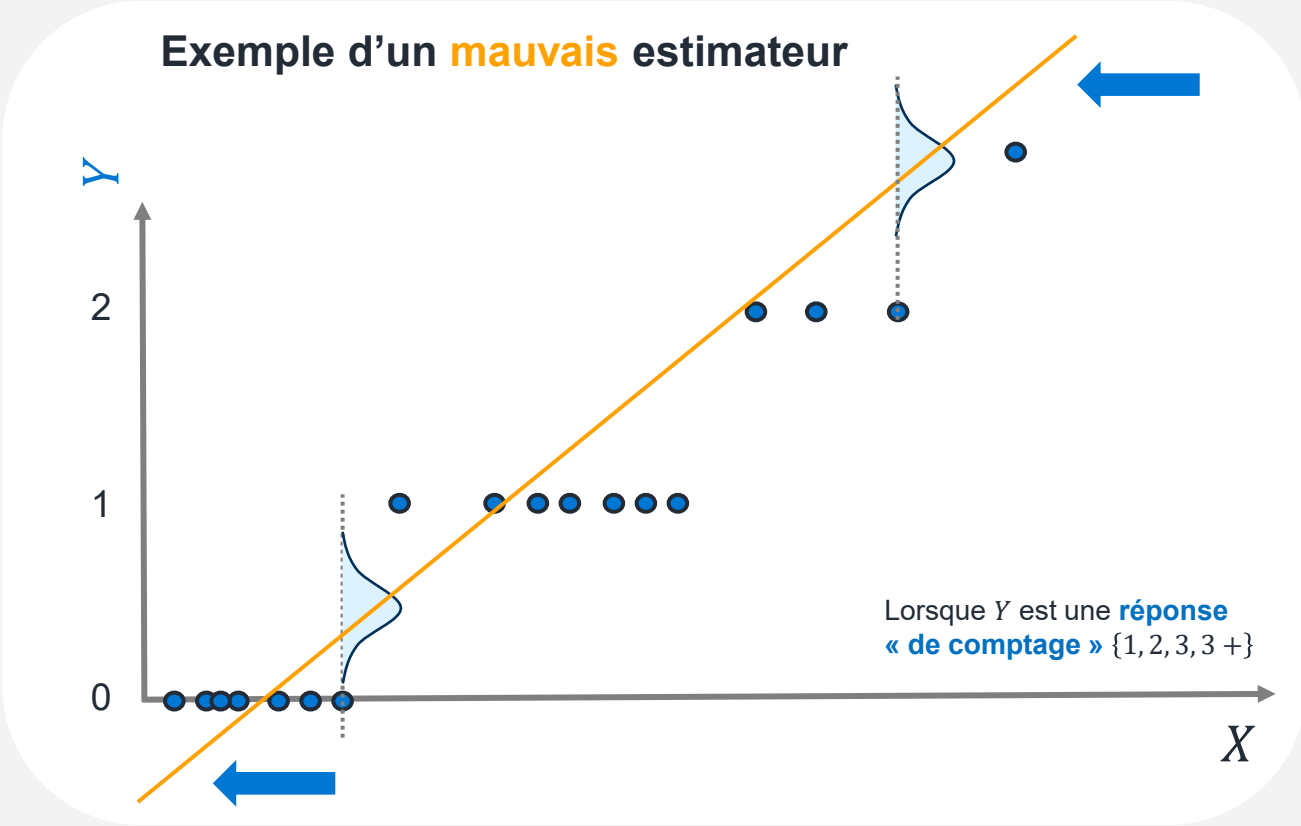
Exemple illustratif

▶ **Modèle linéaire multiple** $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

Lorsque Y n'est pas une variable cible continue alors le **modèle linéaire classique** n'a plus de sens :

1) L'estimateur à pour valeur dans un **ensemble discret**
2) Les **hypothèses sur les résidus** (notamment de variance constante) ne sont pas vérifiées



Limites des modèles linéaires classiques

Cas variable cible **discrète**

Exemple illustratif

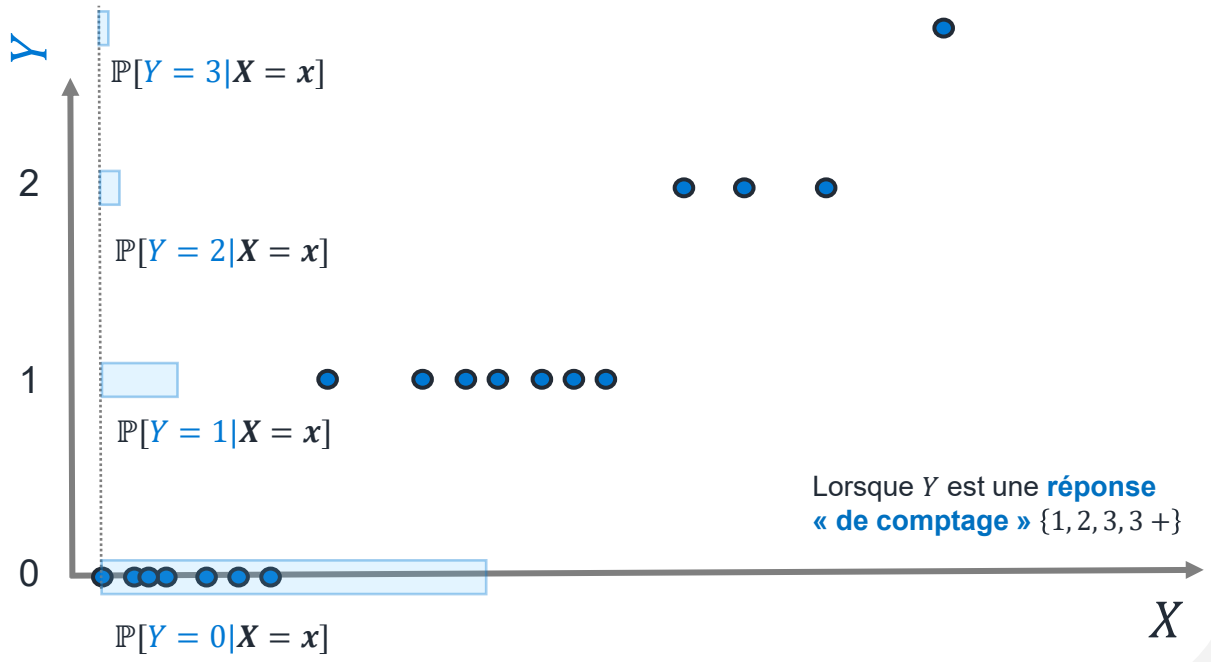
▶ **Solution :** Prédire la **probabilité d'un événement discret** à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$~~

Réponse discrète positive : $\mathbb{P}(Y = k \mid \boldsymbol{X}) =$?

Objectif



Lorsque Y est une **réponse « de comptage »** $\{1, 2, 3, 3+\}$

Limites des modèles linéaires classiques

Cas variable cible **discrète**

Exemple illustratif

▶ **Solution :** Prédire la **probabilité d'un événement discret** à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

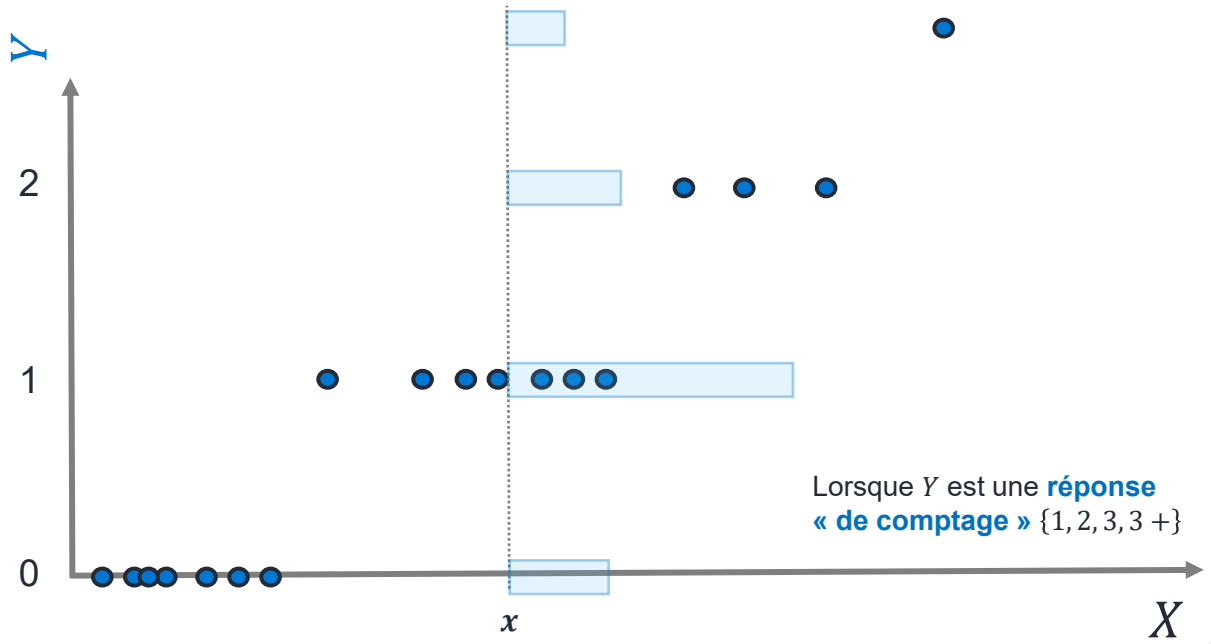
Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$~~

Distribution « naturelle » : $Y \mid \boldsymbol{X} \sim Poisson(\boldsymbol{\lambda})$

Réponse discrète positive : $\mathbb{P}(Y = k \mid \boldsymbol{X}) = \frac{\boldsymbol{\lambda}^k}{k!} e^{-\boldsymbol{\lambda}}$

Lien ? $\boldsymbol{\lambda} \leftrightarrow \boldsymbol{\beta}$

Objectif



Limites des modèles linéaires classiques

Cas variable cible **discrète**

Exemple illustratif

▶ **Solution :** Prédire la **probabilité d'un événement discret** à partir de variables explicatives $\mathbf{X} = \boldsymbol{x}$.

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

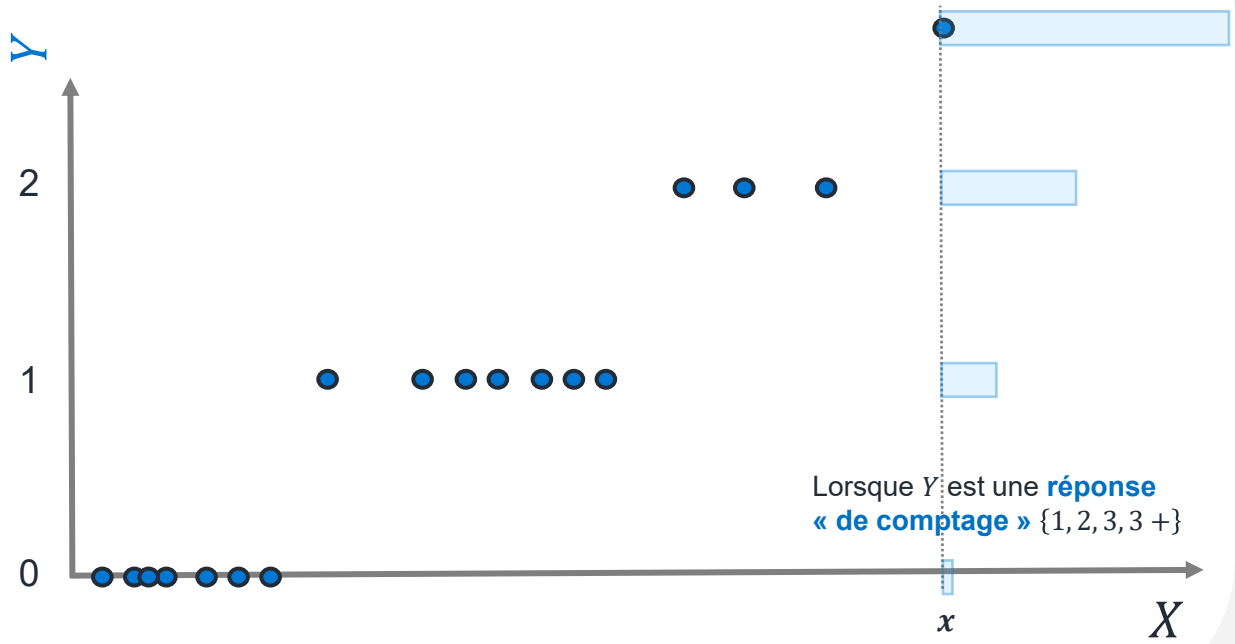
Modèle linéaire : ~~$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{T}}$~~

Distribution « naturelle » : $Y \mid \boldsymbol{X} \sim Poisson(\boldsymbol{\lambda})$

Réponse discrète positive : $\mathbb{P}(Y = k \mid \boldsymbol{X}) = \frac{\boldsymbol{\lambda}^k}{k!} e^{-\boldsymbol{\lambda}}$

Lien ? $\boldsymbol{\lambda} \leftrightarrow \boldsymbol{\beta}$

Objectif



2. Distributions de la famille exponentielle

Milliman

Introduction à la famille exponentielle (1/3)

Loi Normale : Modélisation des données **continues** avec variance **constante**

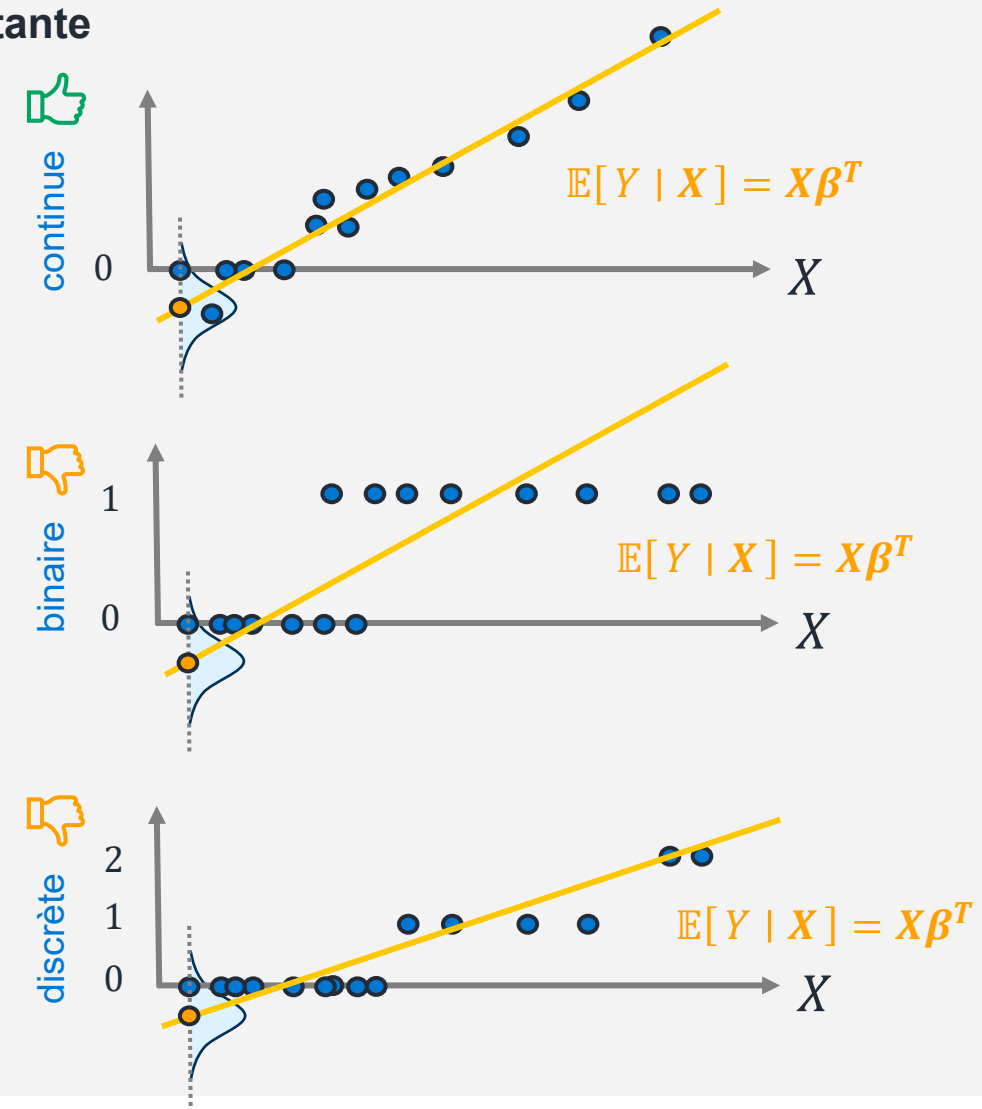
Modèle linéaire multiple

$Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

donc : $Y \mid \boldsymbol{X} \sim N(\boldsymbol{X\beta^T}, \sigma^2 I)$

avec :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \boldsymbol{X\beta^T}}{\sigma}\right)^2\right)$$



Introduction à la famille exponentielle (2/3)

Famille exponentielle : Des données **discrètes** ou continues avec variance constante ou **non constante**

Modèle linéaire multiple

$Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$

donc : $Y \mid \boldsymbol{X} \sim N(\boldsymbol{X\beta^T}, \sigma^2 I)$

avec :

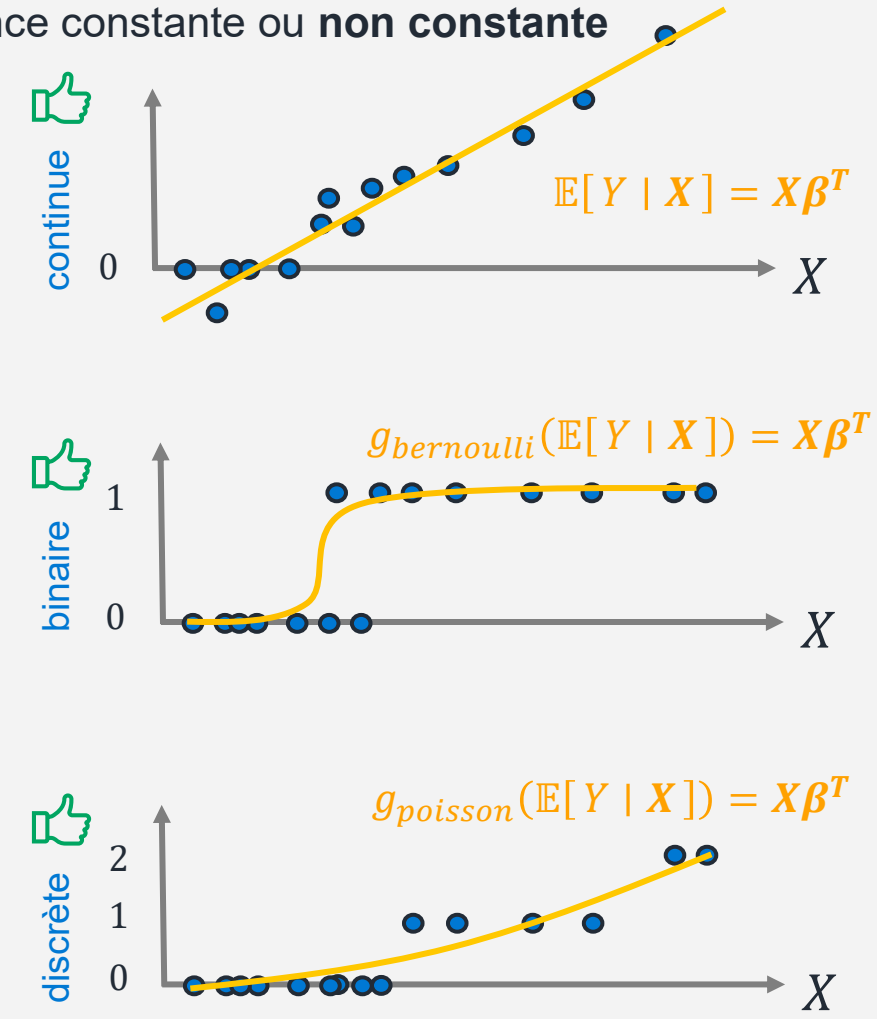
$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \boldsymbol{X\beta^T}}{\sigma}\right)^2\right)$$

- **Extension :** Les **distributions de la famille exponentielle** sont un ensemble de distributions paramétriques :

$$f(y) = c(y, \phi) \exp\left(\frac{y\beta - a(\beta)}{\phi}\right)$$

Incluent des distributions couramment utilisées pour modéliser divers types de données : **continues**, **discrètes**, et **binaires**.

$\mathbb{E}[\, Y \mid \boldsymbol{X} \,] = a'(\beta)$ $\quad$ $\mathrm{Var}[Y|\boldsymbol{X}] = \phi \times a''(\beta)$



Introduction à la famille exponentielle (3/3)

Famille exponentielle : Des données **discrètes** ou continues avec variance constante ou **non constante**

▶ **Distribution de la famille exponentielle** :

$$f(y) = c(y, \phi) \exp\left(\frac{y\beta - a(\beta)}{\phi}\right)$$

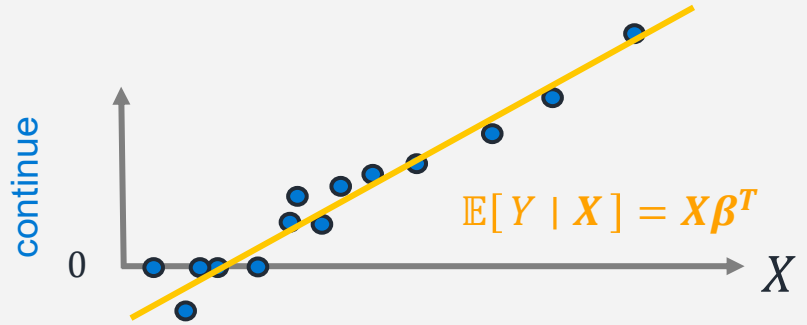
Pour ajuster les données en conséquence, nous devons les transformer à l'aide de fonctions appelées **fonctions de lien**, notées g :

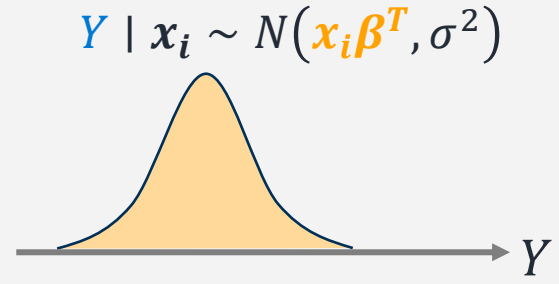
$$g(\mathbb{E}[Y|\boldsymbol{X}]) = \boldsymbol{X\beta^T}$$

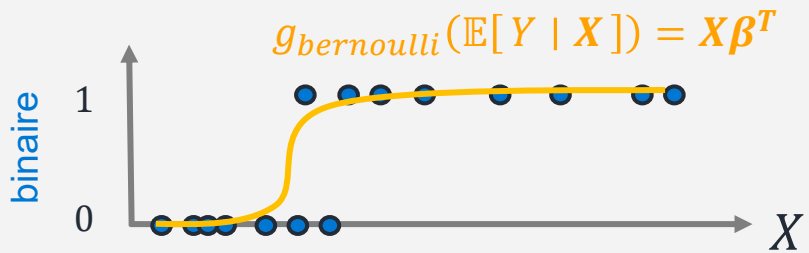
Nous détaillerons dans **la section suivante** les graphiques et les résultats statistiques. Notamment :

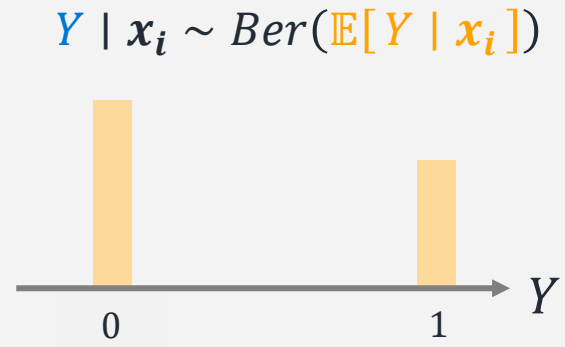
$g_{bernoulli}$?

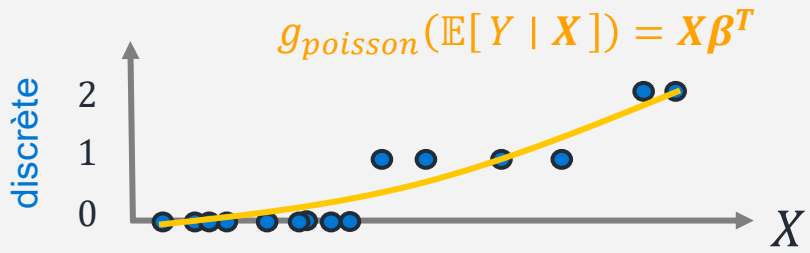
$g_{poisson}$?

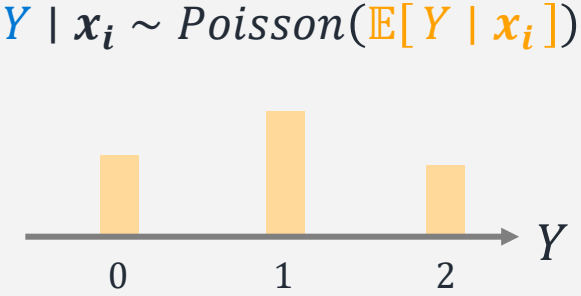












3. Modèles GLM

- **Modèle Logistique**
- Modèle de Poisson
- Récapitulatif

Milliman

Variable cible binaire

Limites des modèles linéaires classiques

Exemple illustratif

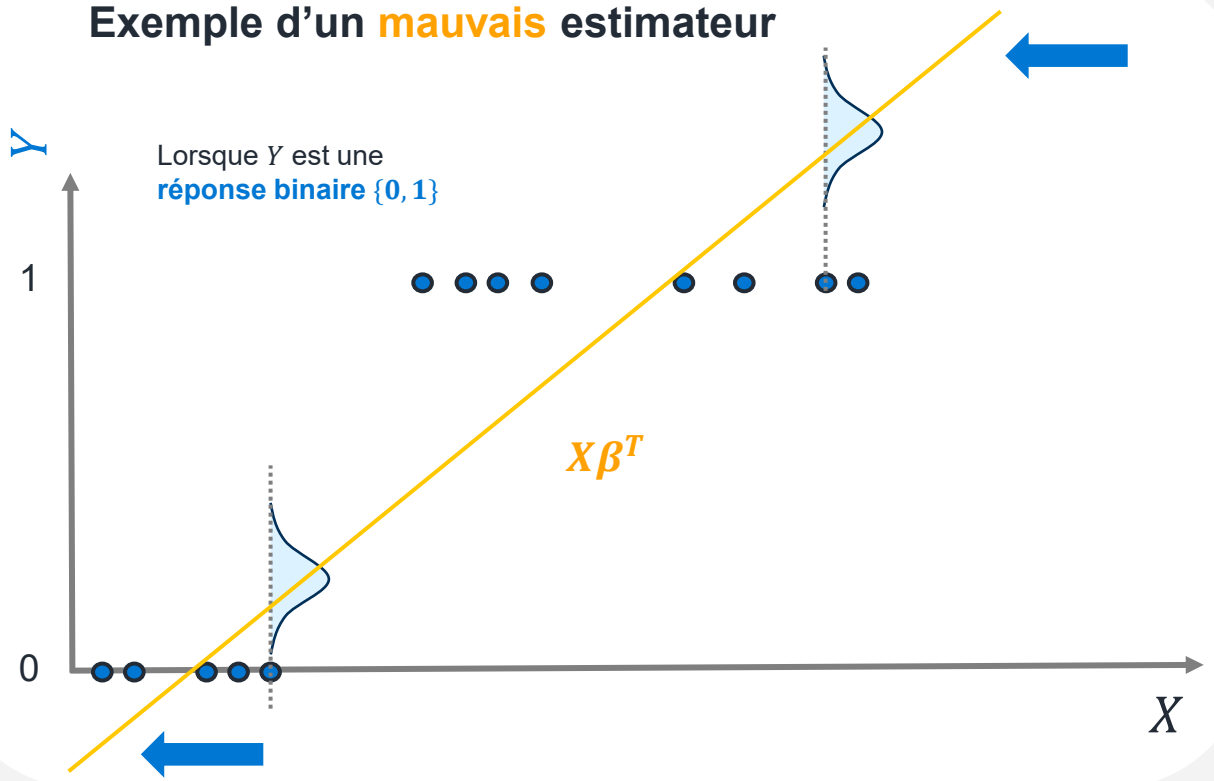
▶ **Modèle linéaire multiple** $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

Lorsque Y est une **réponse binaire** $\{0, 1\}$, nous avons clairement plusieurs problèmes:

- L'estimateur **a pour valeur dans** $\mathbb{R}$
- Les **hypothèses sur les résidus** (notamment de variance constante) ne sont pas vérifiées

Exemple d'un mauvais estimateur



Variable cible binaire

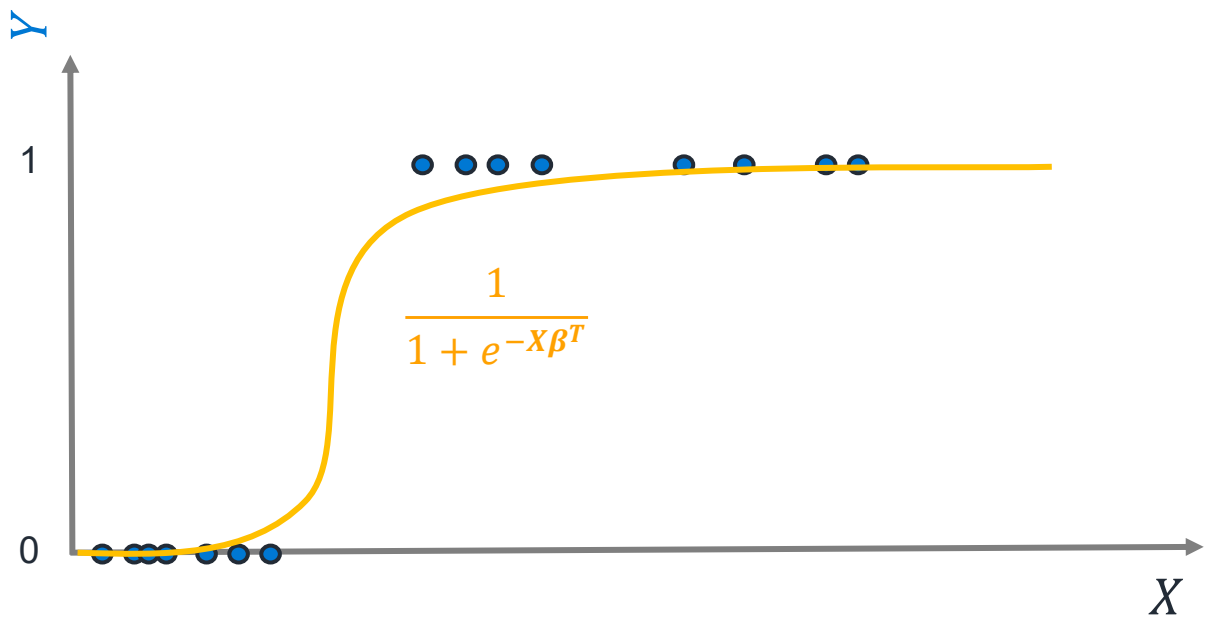
Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

- **Régression Logistique**. Modèle statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

Exemple d'un bon estimateur



Variable cible binaire

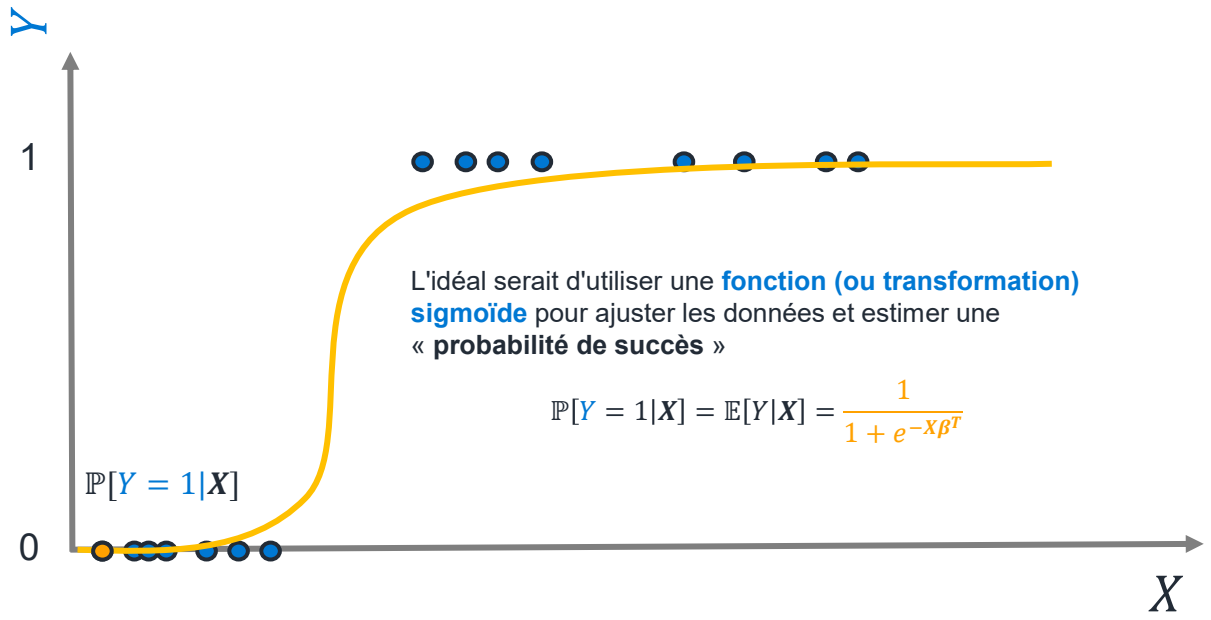
Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu \coloneqq \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 \coloneqq Var[Y|\boldsymbol{X}]$

Coût sinistre	***Occurrence sinistre***	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

- **Régression Logistique**. Modèle statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

Exemple d'un bon estimateur



Variable cible binaire

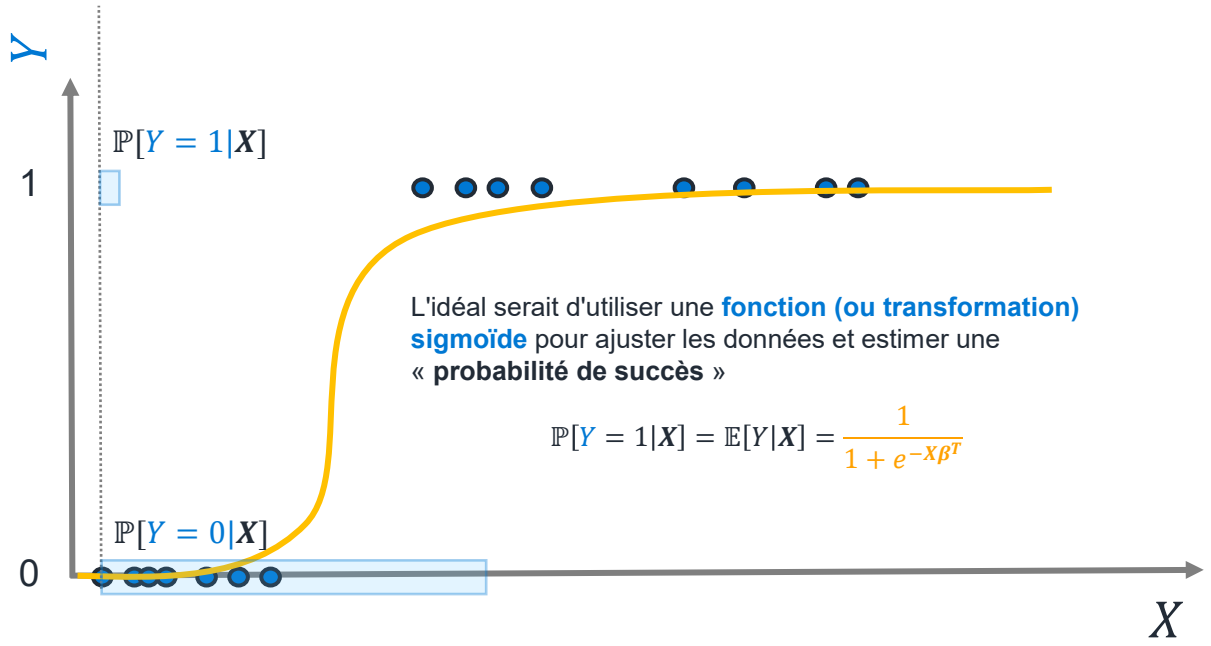
Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

- **Régression Logistique**. Modèle statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

Exemple d'un bon estimateur



Variable cible binaire

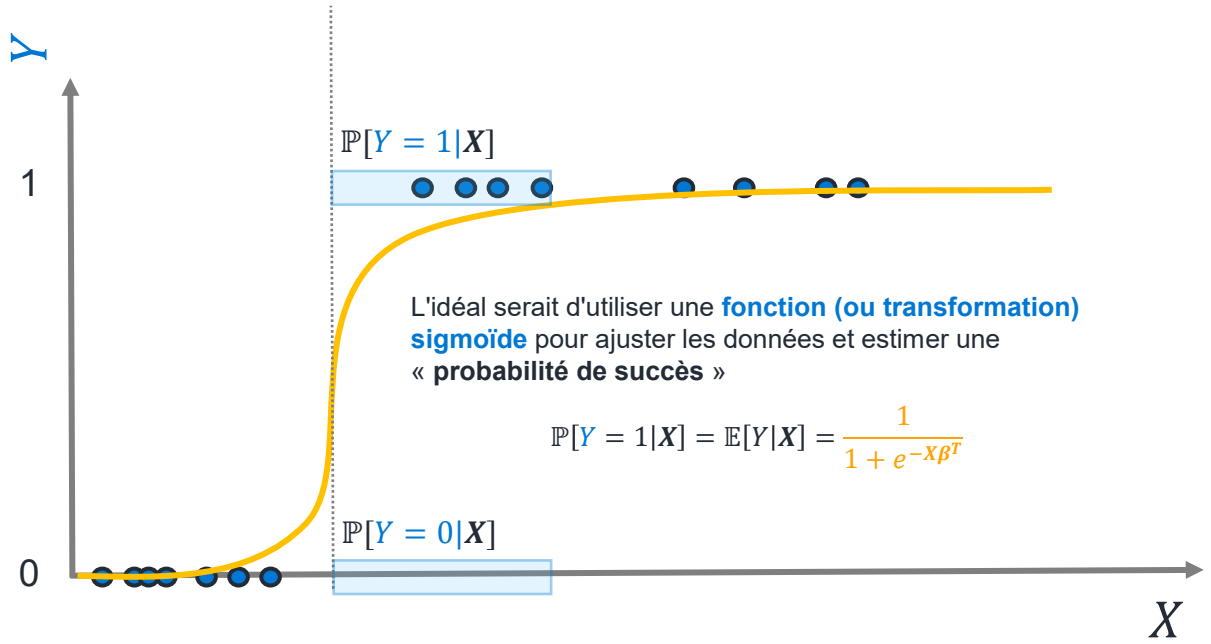
Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

- **Régression Logistique**. Modèle statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

Exemple d'un bon estimateur



Variable cible binaire

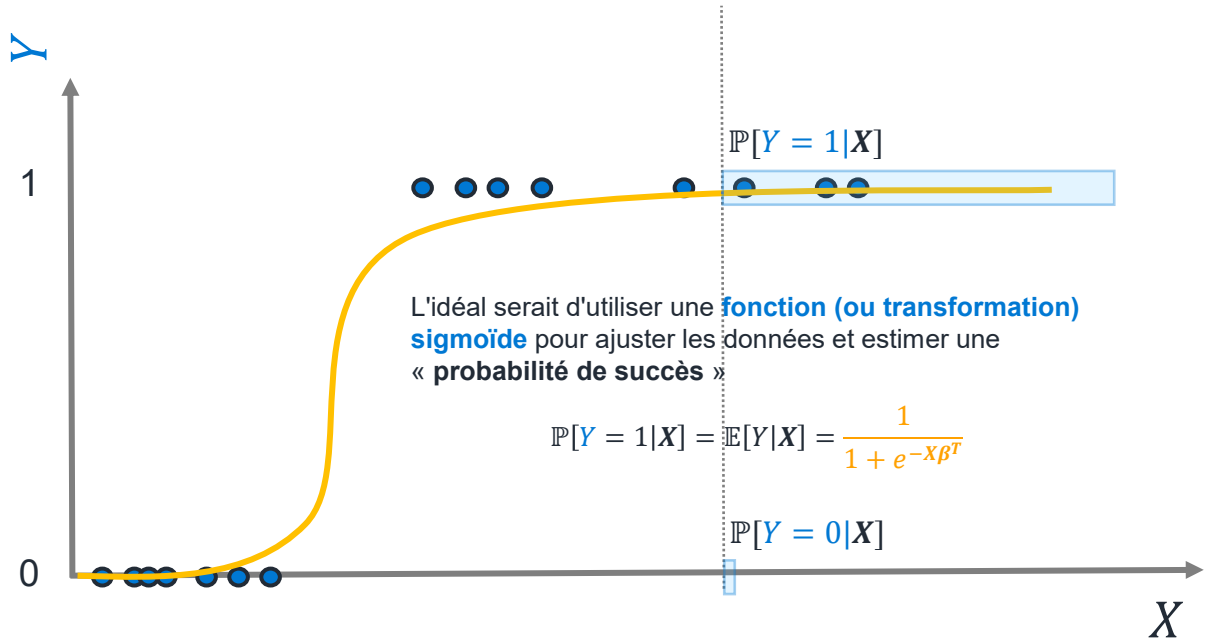
Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	**0**	0
122.3	**1**	1
977.8	**1**	2
⋮	⋮	⋮
1203.1	**1**	3

- **Régression Logistique**. Modèle statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

Exemple d'un bon estimateur



Variable cible binaire

Modèle Logistique

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := Var[Y|\boldsymbol{X}]$

- **Régression Logistique**. Une technique statistique utilisée pour prédire la probabilité d'un événement binaire (0 ou 1) à partir de variables explicatives. Elle est souvent employée pour des tâches de classification.

  Pour ajuster les données, nous devons les transformer à l'aide de fonctions appelées **fonctions de lien**, notées g :

$$g(\mathbb{E}[Y|\boldsymbol{X}]) = \boldsymbol{X\beta^T}$$

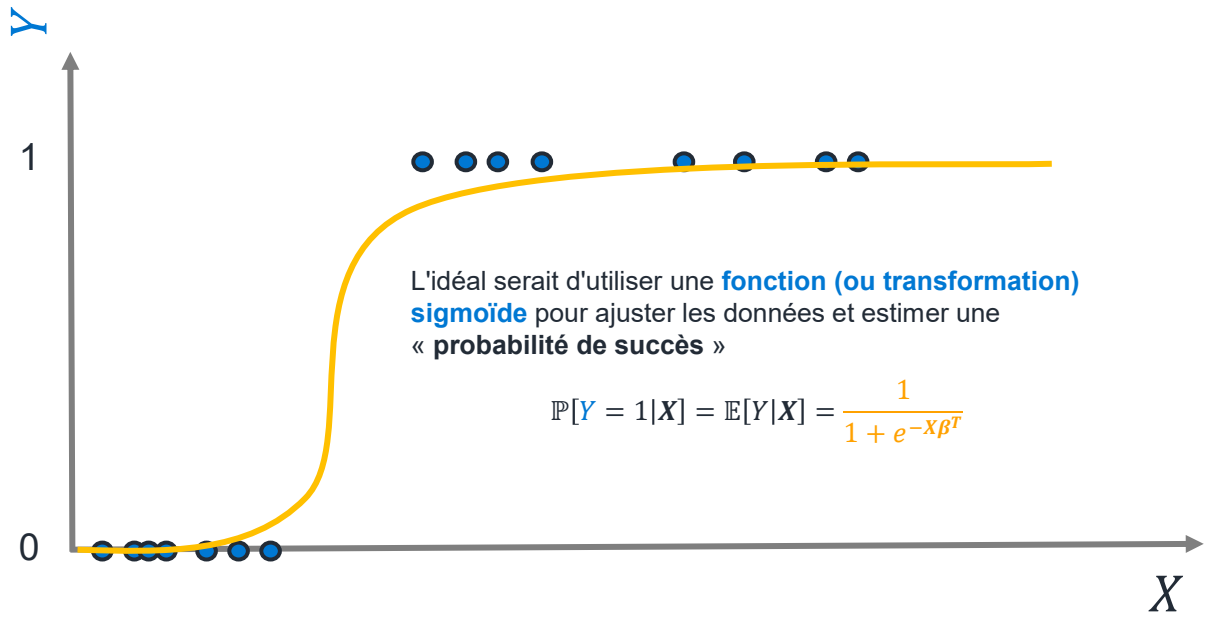
  Fonction logistique comme fonction de lien **:**

$$g(\mathbb{E}[Y|\boldsymbol{X}]) = \log\left(\frac{\mathbb{E}[Y|\boldsymbol{X}]}{1 - \mathbb{E}[Y|\boldsymbol{X}]}\right)$$

  avec

$$\mathbb{E}[Y|\boldsymbol{X}] = \frac{1}{1 + e^{-\boldsymbol{X\beta^T}}}$$

Exemple d'un bon estimateur



3. Modèles GLM

Modèle Logistique

Modèle de Poisson

Récapitulatif

ISFA
Institut de Science Financière et d'Assurances

variable cible de comptage

Cas variable cible **discrète**

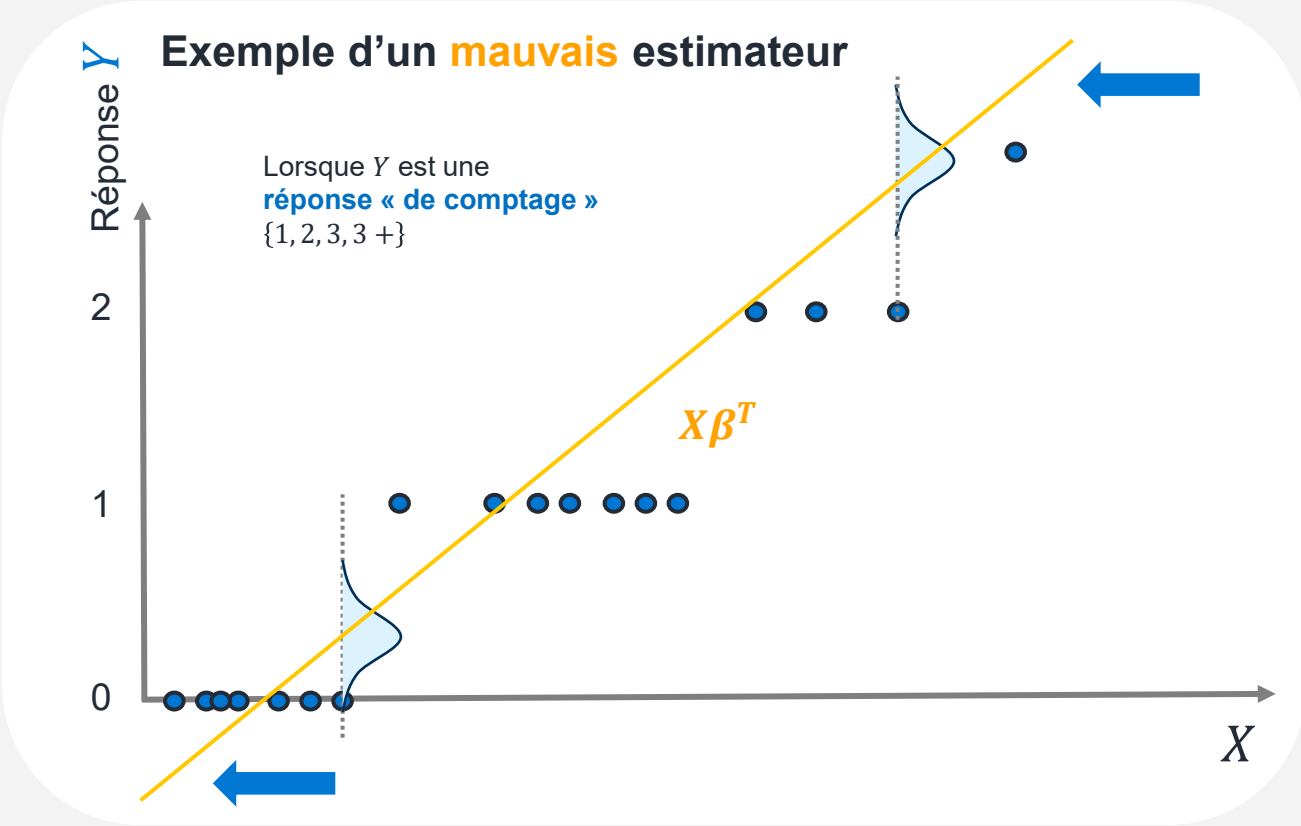
Exemple illustratif

▶ **Modèle linéaire multiple** $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu \coloneqq \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 \coloneqq \mathrm{Var}[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

Lorsque Y est une **réponse discrète (ici, de comptage)** nous avons clairement plusieurs problèmes:

- Les valeurs estimées devraient être des nombres entiers
- Les **hypothèses sur les résidus** (notamment de variance homogène, normalité etc. …) ne sont pas vérifiées



variable cible de comptage

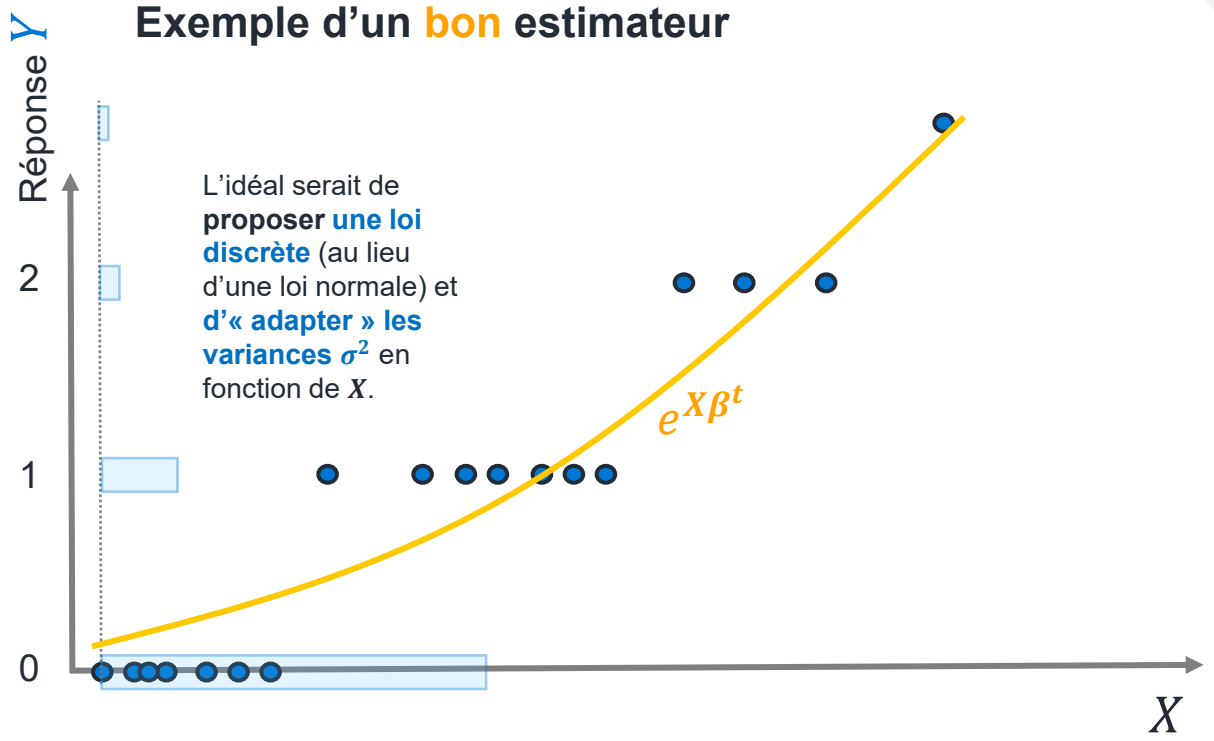
Cas variable cible **discrète**

Exemple illustratif

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := \mathrm{Var}[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

- **Régression de Poisson**. Pour modéliser le nombre de fois qu'un événement se produit dans un intervalle de temps ou d'espace, en fonction de variables explicatives. Elle est particulièrement **adaptée aux données de comptage**.



variable cible de comptage

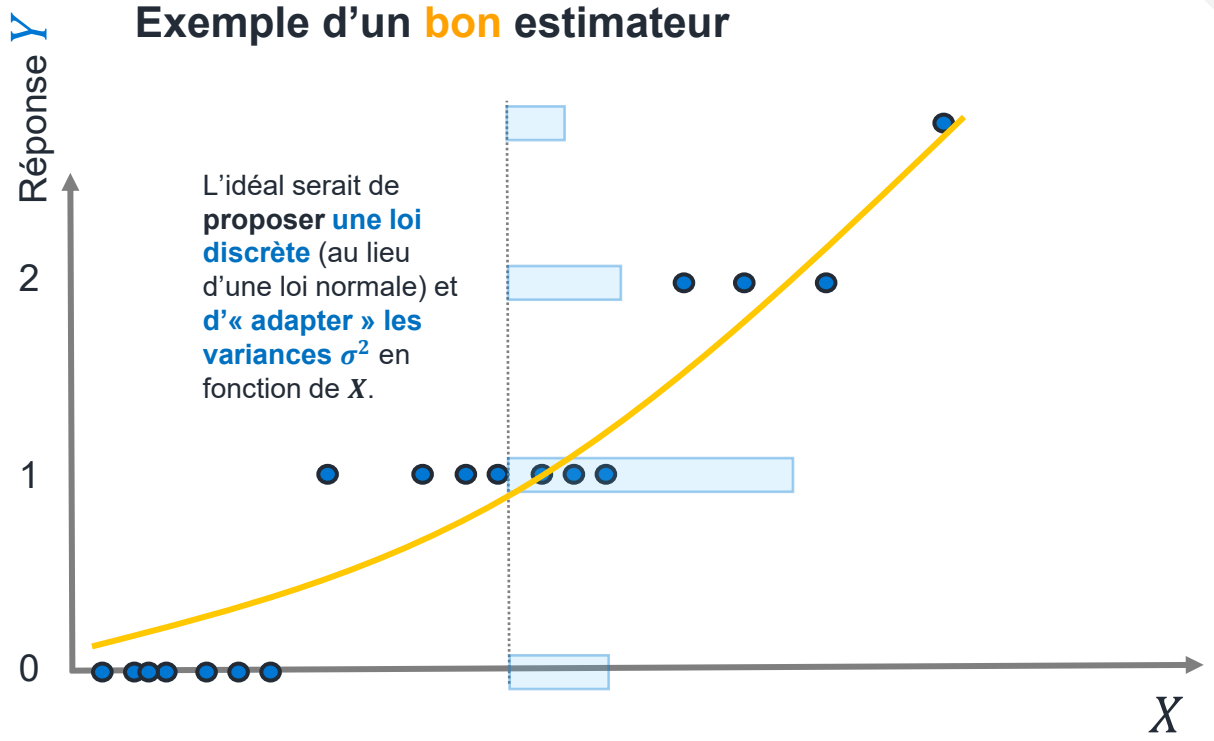
Cas variable cible **discrète**

Exemple illustratif

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := \mathrm{Var}[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

- **Régression de Poisson**. Pour modéliser le nombre de fois qu'un événement se produit dans un intervalle de temps ou d'espace, en fonction de variables explicatives. Elle est particulièrement **adaptée aux données de comptage**.



variable cible de comptage

Cas variable cible **discrète**

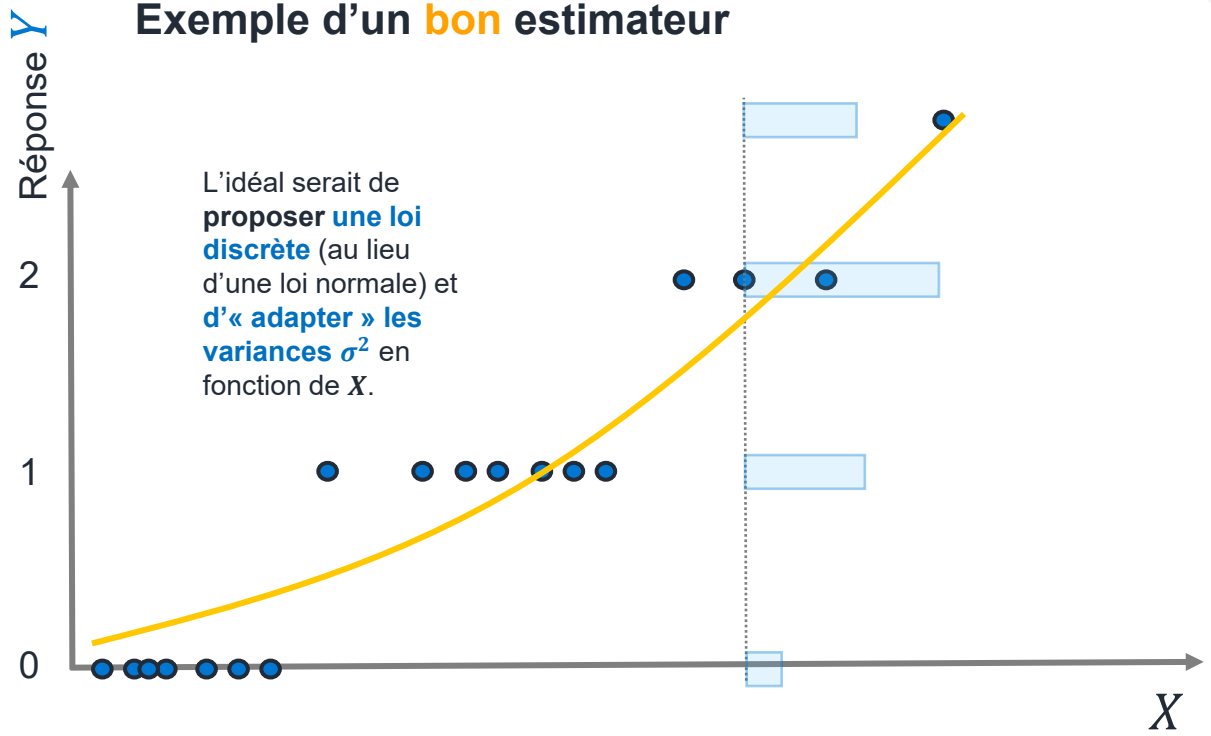
Exemple illustratif

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := \mathrm{Var}[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

- **Régression de Poisson**. Pour modéliser le nombre de fois qu'un événement se produit dans un intervalle de temps ou d'espace, en fonction de variables explicatives. Elle est particulièrement **adaptée aux données de comptage**.

Exemple d'un bon estimateur



variable cible de comptage

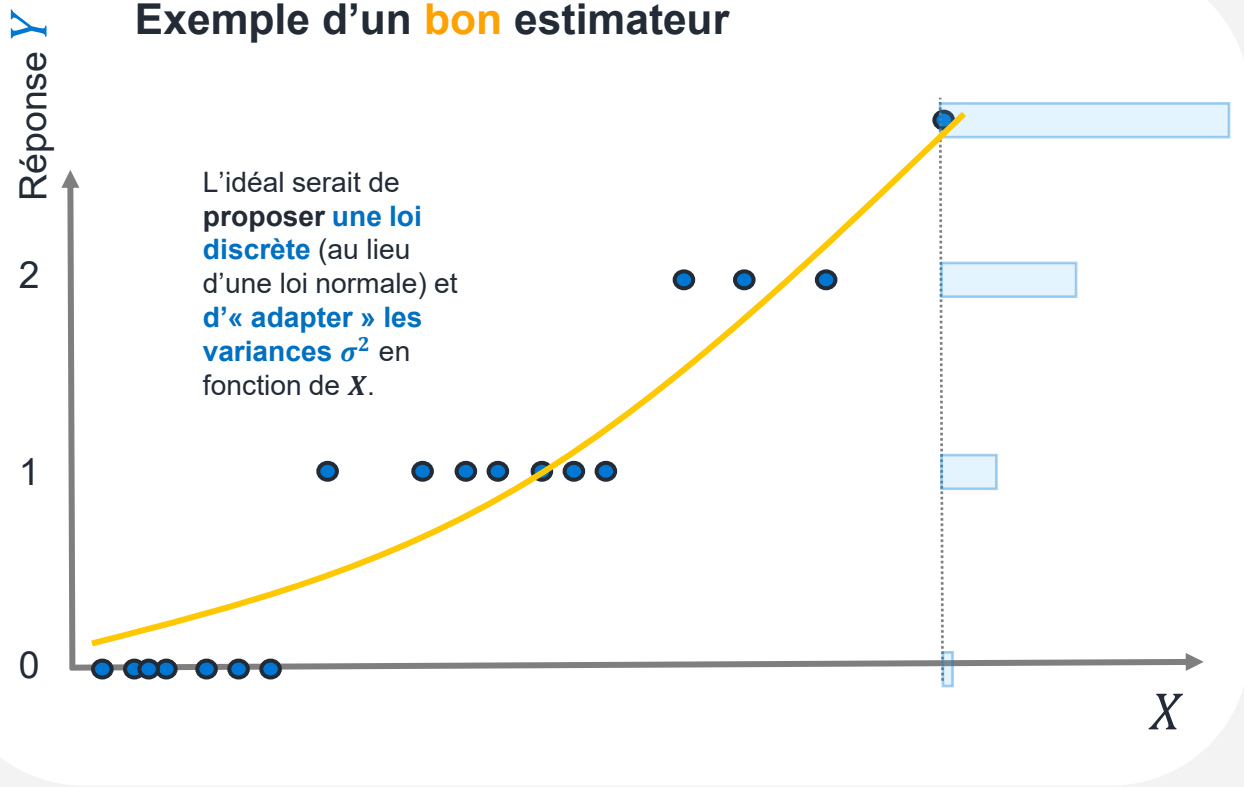
Cas variable cible **discrète**

Exemple illustratif

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := \mathrm{Var}[Y|\boldsymbol{X}]$

Coût sinistre	*Occurrence sinistre*	*Nombre sinistres*
0	0	**0**
122.3	1	**1**
977.8	1	**2**
⋮	⋮	⋮
1203.1	1	**3**

- **Régression de Poisson**. Pour modéliser le nombre de fois qu'un événement se produit dans un intervalle de temps ou d'espace, en fonction de variables explicatives. Elle est particulièrement **adaptée aux données de comptage**.



variable cible de comptage

Modèle de Poisson

Exemple illustratif

Modèle linéaire multiple $Y = \boldsymbol{X\beta^T} + \boldsymbol{\varepsilon}$ avec $\varepsilon_i \; iid \sim N(0, \sigma^2)$ $\mu := \mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$ et $\sigma^2 := \mathrm{Var}[Y|\boldsymbol{X}]$

- **Régression de Poisson**. Une méthode statistique utilisée pour modéliser le nombre de fois qu'un événement se produit dans un intervalle de temps ou d'espace, en fonction de variables explicatives. Elle est particulièrement adaptée aux données de comptage:

$Y \mid \boldsymbol{X} \sim Pois(\lambda)$ avec $\lambda = \mathbb{E}[\,Y \mid \boldsymbol{X}\,] = e^{\boldsymbol{X\beta^t}}$ et

$$\mathbb{P}(Y = k \mid \boldsymbol{X}) = \frac{\lambda^k}{k!} e^{-k}$$

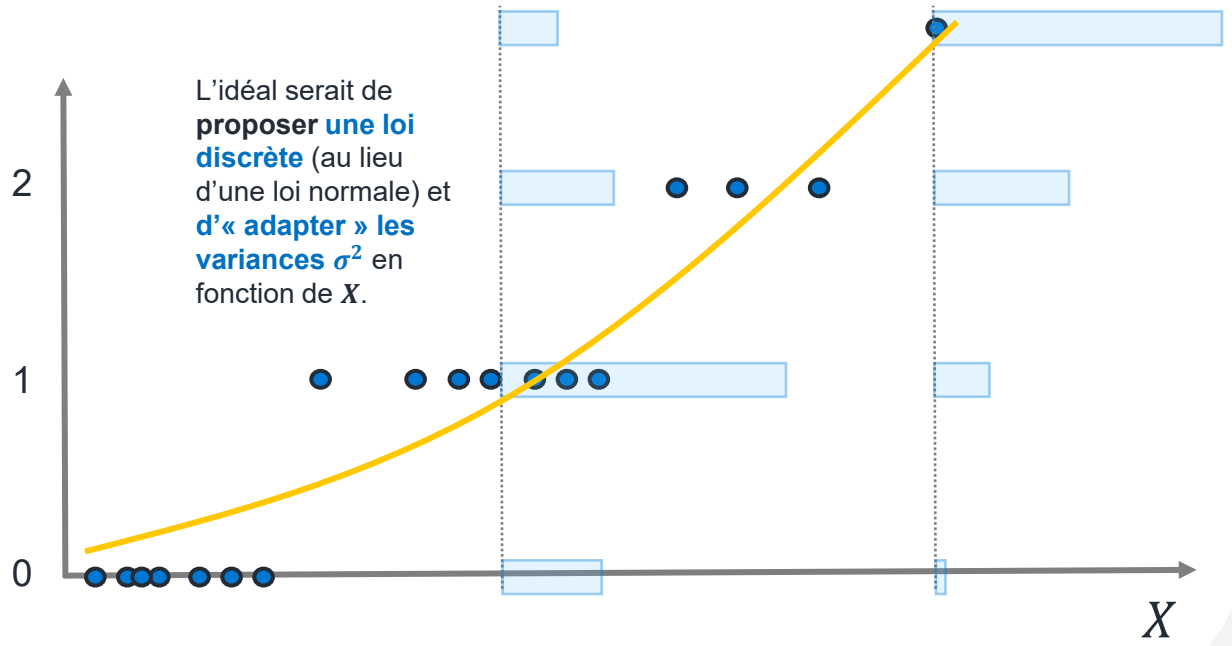
La **fonction logarithmique** comme fonction de lien :

$$g(\mathbb{E}[Y|\boldsymbol{X}]) = \log(\mathbb{E}[Y|\boldsymbol{X}]) = \boldsymbol{X\beta^T}$$

avec

$$\mathbb{E}[Y|\boldsymbol{X}] = e^{\boldsymbol{X\beta^t}}$$

Exemple d'un bon estimateur



3. Modèles GLM

Modèle Logistique

Modèle de Poisson

Récapitulatif

ISFA
Institut de Science Financière et d'Assurances

Formulation générale

Modèles linéaires généralisés

Modèle linéaire généralisé ou **GLM** (pour *Generalized Linear Model*) étend le modèle linéaire « classique » en :

1) utilisant une **fonction de lien (inversible)** pour relier la variable de réponse au modèle

$$\mathbb{E}[Y|\boldsymbol{X}] = g^{-1}(\boldsymbol{X\beta^T})$$

2) et en permettant à la **variance de dépendre de la valeur prédite** selon la distribution choisie (V une fonction)

$$\text{Var}[Y|\boldsymbol{X}] = \phi \times V\left(g^{-1}(\boldsymbol{X\beta^T})\right)$$

avec V une fonction variance qui dépend $\mathbb{E}[Y \mid X]$ et ϕ un paramètre d'échelle.

Modèle classique

Distribution des erreurs : **Normale**

$$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X\beta^T}$$

$$\text{Var}[Y|\boldsymbol{X}] = \sigma^2$$

Modèle logistique

Distribution : **Bernoulli** (ou **Binomiale**)

$$\mathbb{E}[Y|\boldsymbol{X}] = \frac{1}{1+e^{-\boldsymbol{X\beta^T}}}$$

$$\text{Var}[Y|\boldsymbol{X}] = \mu(1-\mu)$$

Modèle de Poisson

Distribution des erreurs : **Poisson**

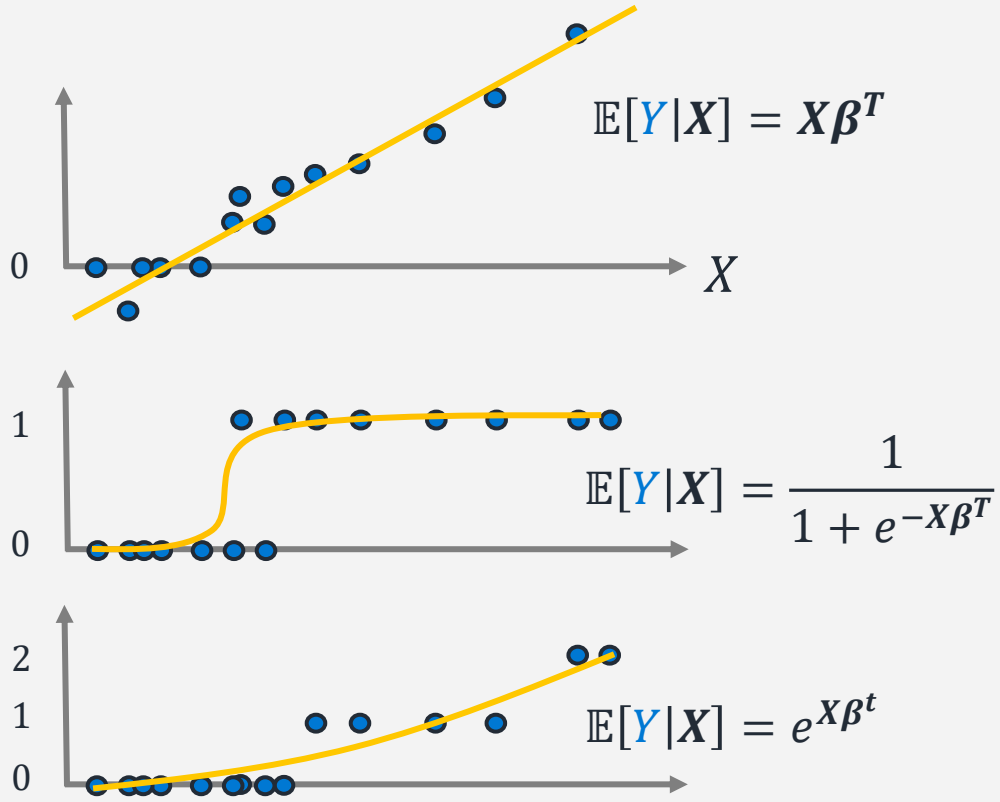
$$\mathbb{E}[Y|\boldsymbol{X}] = e^{\boldsymbol{X\beta^t}}$$

$$\text{Var}[Y|\boldsymbol{X}] = \mathbb{E}[Y \mid \boldsymbol{X}] = \mu$$

Modèles GLM et applications en assurance

Récapitulatif

	Modèle Linéaire « classique »	Modèle Logistique	Modèle de Poisson
Variable continue **(coût sinistre)**	✓		
Variable binaire **(sinistre oui/non)**		✓	
Variable discrète **(nombre** ou **fréquence)**			✓



4. Estimation des paramètres : Maximum de vraisemblance

Fonction de vraisemblance (1/2)

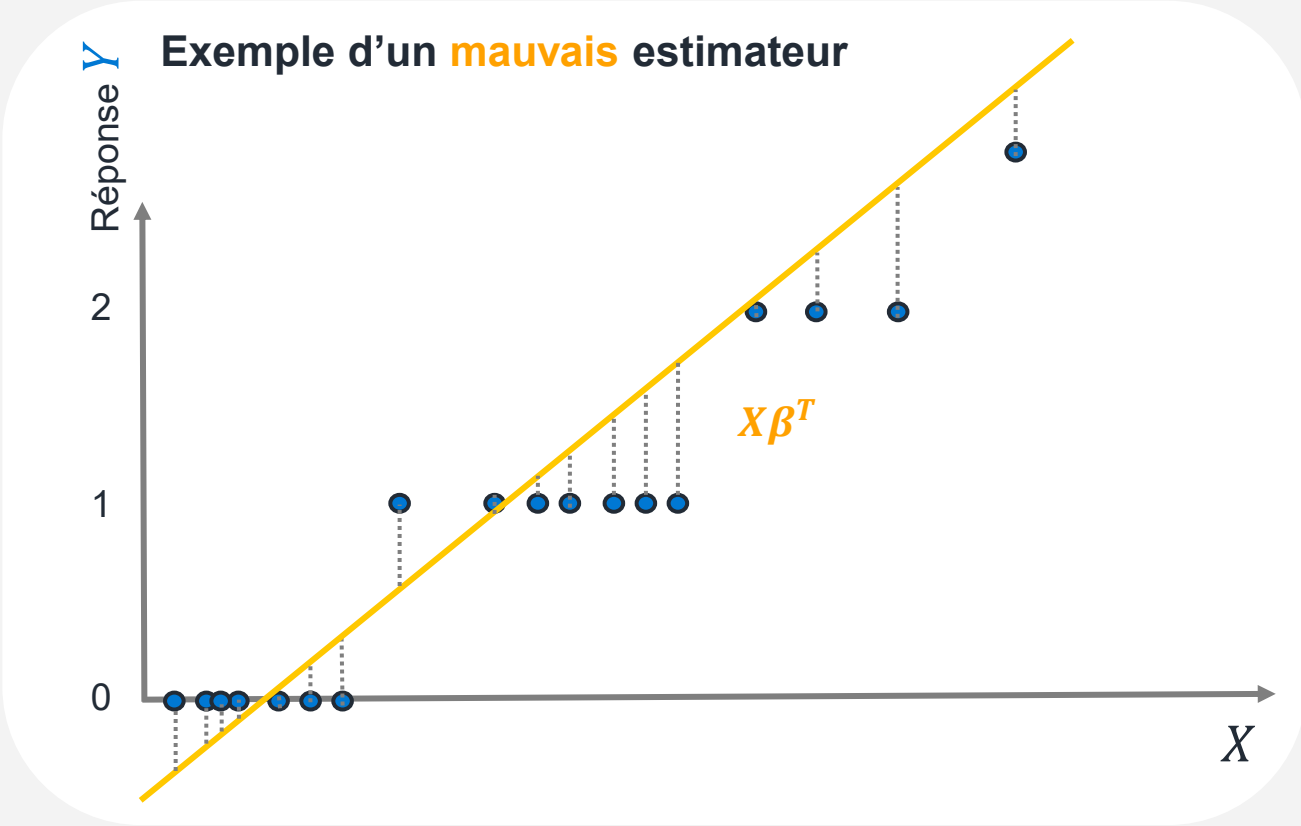
Les limites de la méthode des moindres carrés

Exemple illustratif

La méthode des moindres carrés

$$R(\boldsymbol{\beta}) \coloneqq \sum_i \left(Y_i - \widehat{Y}_i\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

Adaptée **aux relations linéaires** entre les variables



Fonction de vraisemblance (1/2)

Les limites de la méthode des moindres carrés

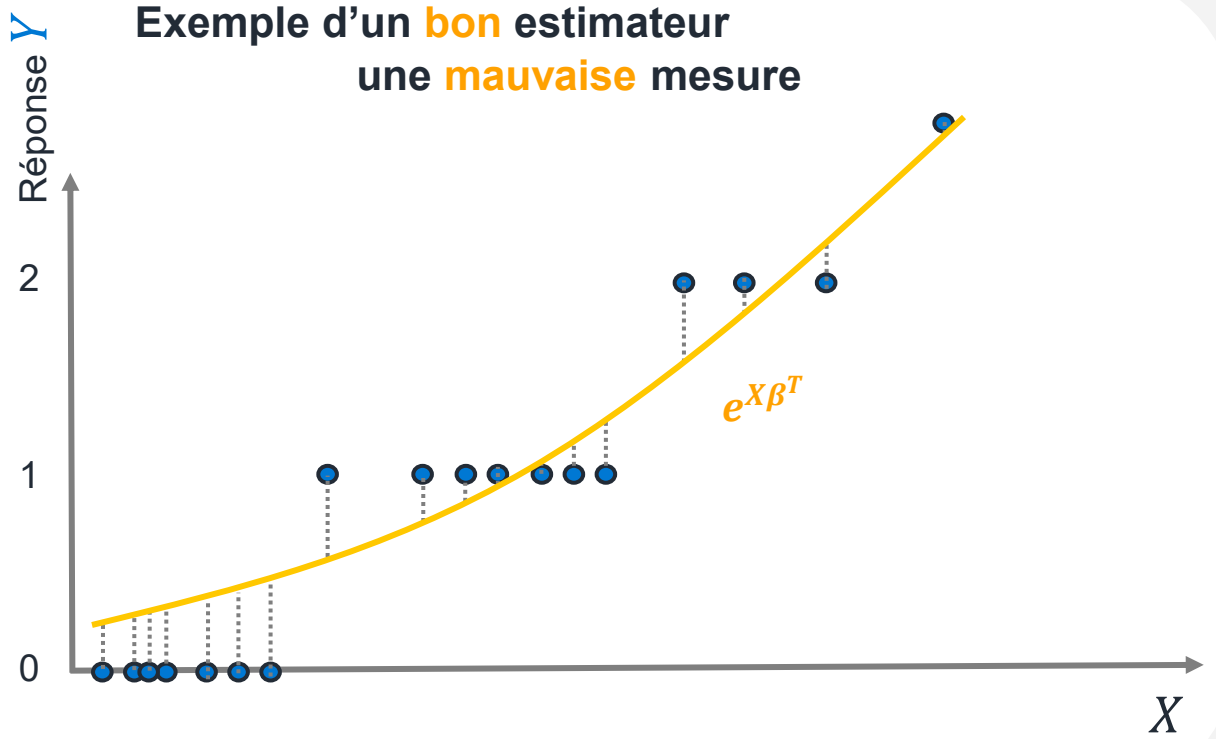
Exemple illustratif

La méthode des moindres carrés

$$R(\boldsymbol{\beta}) \coloneqq \sum_i \left(Y_i - \widehat{Y}_i\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

Adaptée **aux relations linéaires** entre les variables

En présence d'une **relation non linéaire**, son utilisation peut entraîner des erreurs significatives



Fonction de vraisemblance (2/2)

Les limites de la méthode des moindres carrés

Exemple illustratif

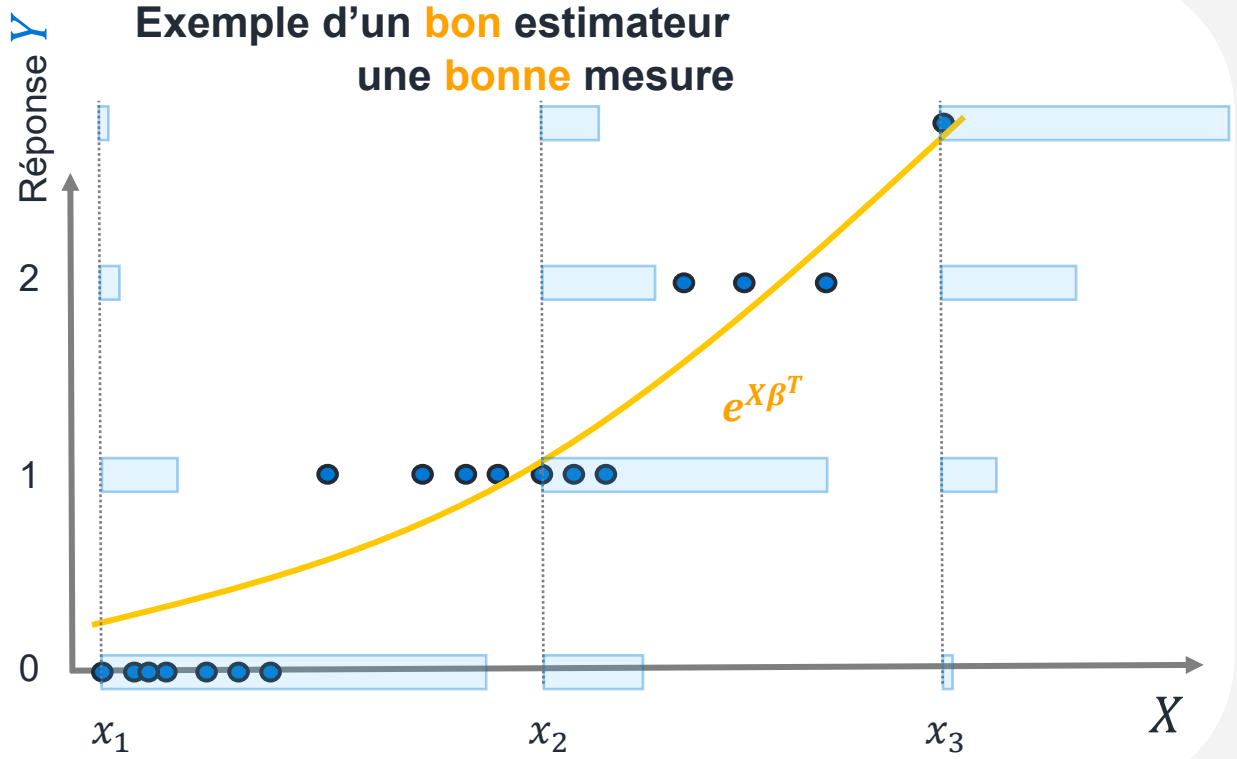
La méthode des moindres carrés

$$R(\boldsymbol{\beta}) \coloneqq \sum_i \left(Y_i - \widehat{Y}_i\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

- Adaptée **aux relations linéaires** entre les variables
- En présence d'une **relation non linéaire**, son utilisation peut entraîner des erreurs significatives

- Estimer les paramètres de **la fonction de densité / probabilité**

$$f_{\boldsymbol{\beta}}(y_i) \coloneqq \mathbb{P}_{\boldsymbol{\beta}}(Y_i = y_i \mid \boldsymbol{X})$$



Fonction de vraisemblance (2/2)

Les limites de la méthode des moindres carrés

Exemple illustratif

La méthode des moindres carrés

$$R(\boldsymbol{\beta}) \coloneqq \sum_i \left(Y_i - \widehat{Y_i}\right)^2 = \sum_i \hat{\varepsilon}_i^2$$

Adaptée **aux relations linéaires** entre les variables

En présence d'une **relation non linéaire**, son utilisation peut entraîner des erreurs significatives

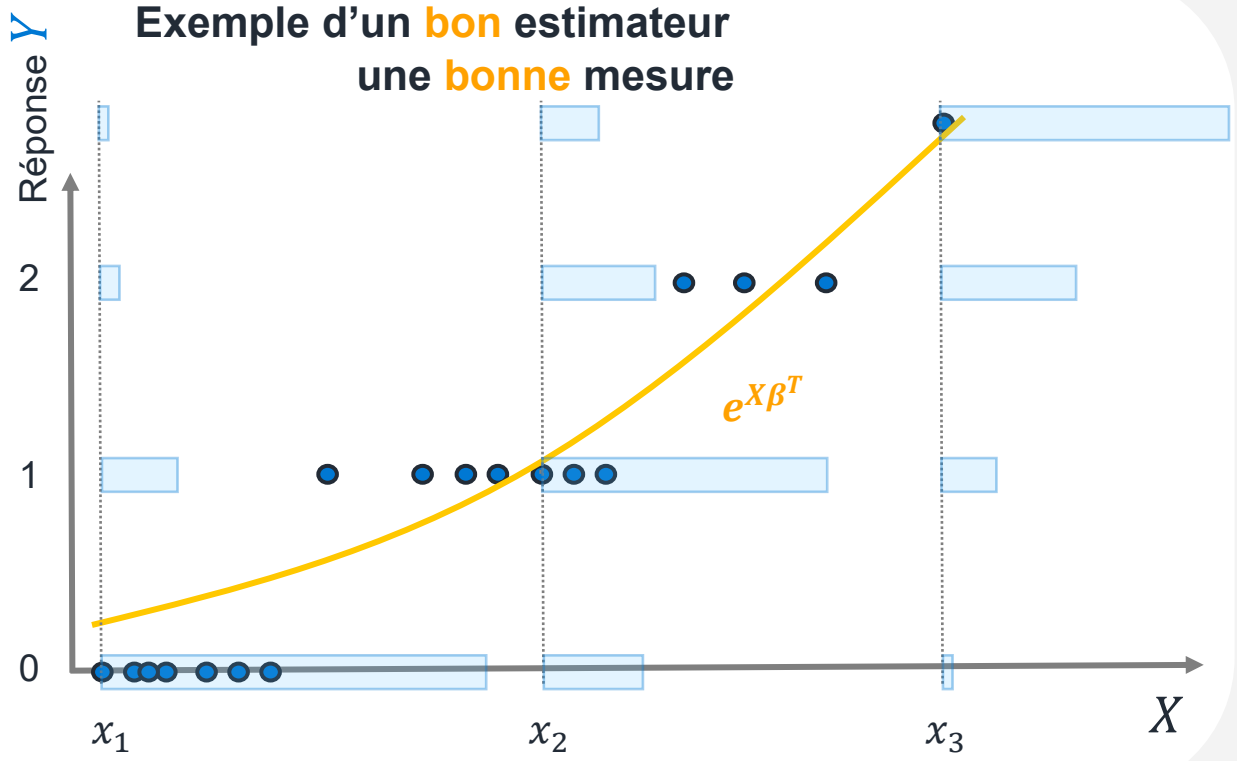
Estimer les paramètres de **la fonction de densité / probabilité**

$$f_{\boldsymbol{\beta}}(y_i) \coloneqq \mathbb{P}_{\boldsymbol{\beta}}(Y_i = y_i \mid \boldsymbol{X})$$

- **Fonction de vraisemblance** (ou **log de vraisemblance**)

$$L(\boldsymbol{y}, \boldsymbol{\beta}) \coloneqq \prod_{i=1,\dots,n} f_{\boldsymbol{\beta}}(y_i)$$

$$l(\boldsymbol{y}, \boldsymbol{\beta}) \coloneqq \log L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$



Maximum de vraisemblance

Cas modèle de Poisson

Exemple illustratif

Estimateur du maximum de vraisemblance (*MLE* en anglais) : estimateur statistique utilisé pour inférer les **paramètres de la loi de probabilité** d'un échantillon donné en recherchant les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

$$\boldsymbol{\beta}_{GLM} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

Modèle de Poisson

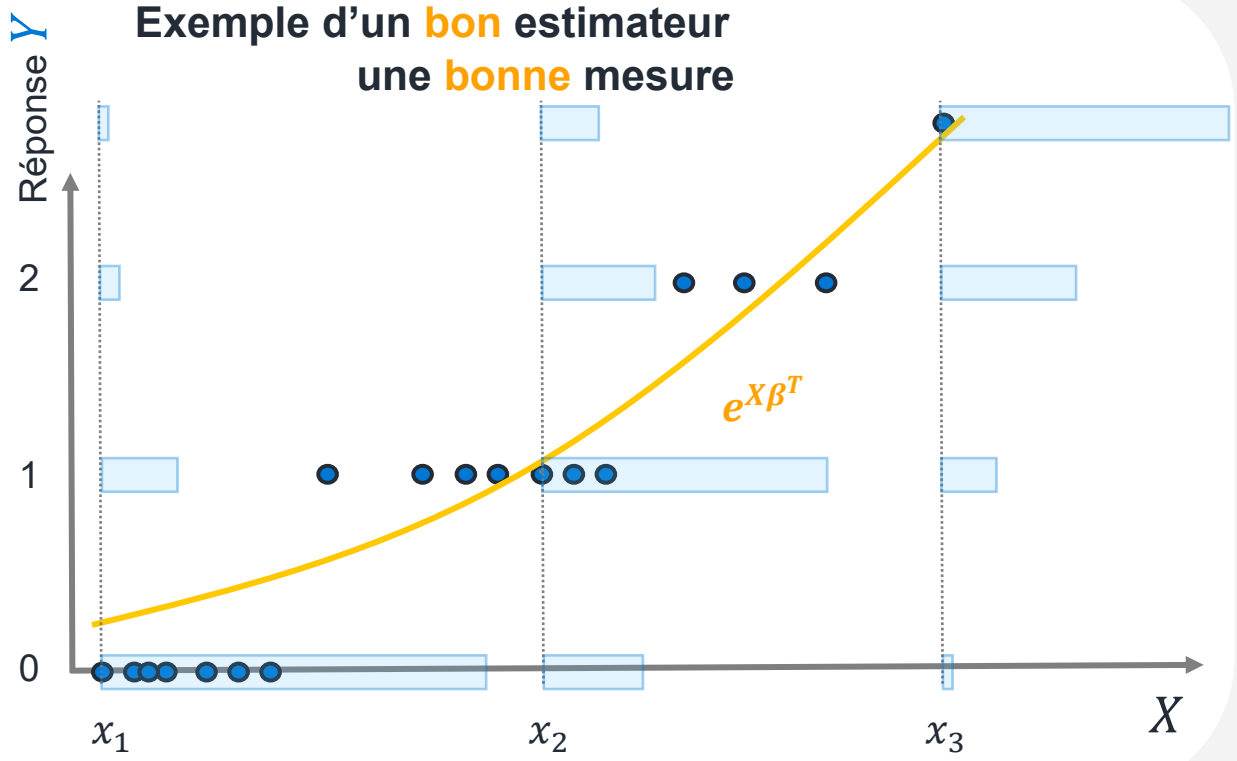
$$\sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i) = \sum_{i=1,\dots,n} \log \frac{\left(\boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{T}}\right)^{y_i}}{y_i!} e^{-y_i}$$

$$= \sum_{i=1,\dots,n} \left(y_i \boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{T}} - e^{\boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{T}}} - \log y_i!\right)$$

Comme nous cherchons $\boldsymbol{\beta}$ qui maximise $l(y, \boldsymbol{\beta})$, nous prenons la simplification suivante :

$$\boldsymbol{\beta}_{GLM} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(y_i \boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{T}} - e^{\boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{T}}}\right)$$

Solution numérique : Descente de gradient



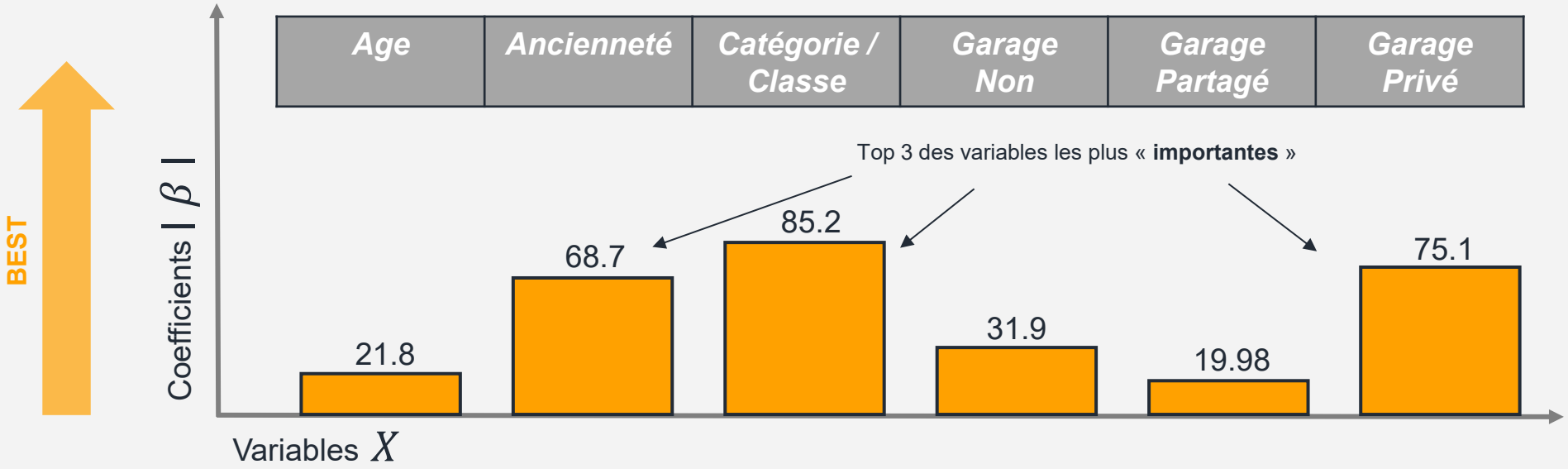
5. Interprétation des coefficients / paramètres

ISFA
Institut de Science Financière et d'Assurances

Importance des variables explicatives (1/2)

Des approches « simples » basées sur les coefficients

Exemple illustratif



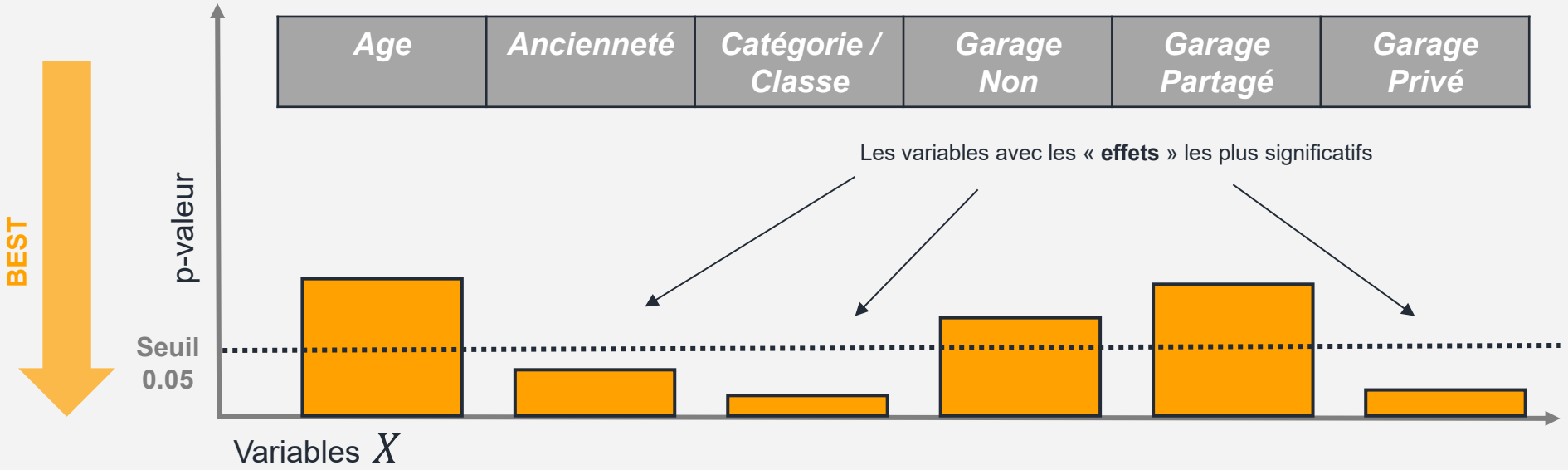
Approche 1 : magnitude des coefficients

Les coefficients estimés mesurent l'effet d'une **variation unitaire d'une variable explicative sur la variable cible**, en maintenant constantes les autres variables. Dans certains cas, les coefficients peuvent être utilisés pour évaluer l'importance relative des variables. **Plus un coefficient est grand (en valeur absolue),** plus l'impact de la variable correspondante est important.

Importance des variables explicatives (2/2)

Des approches « simples » basées sur les coefficients

Exemple illustratif



Approche 1 : magnitude des coefficients

Les coefficients estimés mesurent l'effet d'une **variation unitaire d'une variable explicative sur la variable cible**, en maintenant constantes les autres variables. Dans certains cas, les coefficients peuvent être utilisés pour évaluer l'importance relative des variables. **Plus un coefficient est grand (en valeur absolue),** plus l'impact de la variable correspondante est important.

Approche 2 : signification statistique (p-valeur)

Une autre manière d'évaluer l'importance d'une variable est de tester si son coefficient est statistiquement différent de zéro. Si la **p-valeur** associée au coefficient est inférieure à un seuil conventionnel (**souvent 0.05**), la variable est considérée comme ayant un effet significatif sur la variable cible.

Nous détaillerons plus tard les tests statistiques

Zoom sur le modèle Logistique

Cas assurance auto

Exemple illustratif

~~Coût sinistre~~	Occurrence sinistre	Age	Ancienneté	Catégorie / Classe	Garage Non	Garage Partagé	Garage Privé
~~0~~	0	0.84	0	0.666	0	0	1
~~122.3~~	1	0	1	0.333	0	0	1
~~977.8~~	1	0.04	0.875	0.333	1	0	0
~~0~~	0	1	0.625	0.333	0	1	0
~~311.1~~	1	0.36	0.25	1	1	0	0

Tâche de classification avec une **variable cible binaire**

Les cotes (ou *odds*) pour X_i **:**

$$e^{X\beta^T} = \frac{\mathbb{P}(Y=1 \mid X_i)}{1-\mathbb{P}(Y=1 \mid X_i)} = \frac{\mathbb{P}(Y=1 \mid X_i)}{\mathbb{P}(Y=0 \mid X_i)}$$

- **Rapport des cotes** (*Odds-Ratio ou OR*) pour des **prédicteurs binaires** : C'est le rapport des cotes des probabilités que ce soit positif pour ceux qui ont la caractéristique X_i d'une part et de ceux qui ne l'ont pas d'autre part. Exemple avec

$$OR_i = \frac{\frac{\mathbb{P}(Y=1 \mid X_i=1)}{\mathbb{P}(Y=0 \mid X_i=1)}}{\frac{\mathbb{P}(Y=1 \mid X_i=0)}{\mathbb{P}(Y=0 \mid X_i=0)}} = \exp(\beta_i)$$

- $OR_i = 1$, la variable cible est indépendantes de X_i
- $OR_i > 1$, la variable cible est plus fréquente pour les individus **qui ont** X_i
- $OR_i < 1$, la variable cible est plus fréquente pour les individus **qui n'ont pas** X_i

6. Evaluation des GLM

ISFA
Institut de Science Financière et d'Assurances

Rappel

Matrice de confusion et AUC

Occurrence sinistre	*Score*	*Prédiction (seuil 0.5)*	*Prédiction (seuil 1)*	*Prédiction (seuil 0.65)*
0	0.4	0	0	0
1	0.7	1	0	1
1	0.35	0	0	0
0	0.52	1	0	0
1	0.68	1	0	1
0	0.2	0	0	0
0	0.1	0	0	0
0	0.6	1	0	0
1	0.7	1	0	1
0	0.55	1	0	0
		Accuracy *: 0.6*	***Accuracy*** *: 0.6*	***Accuracy*** *: 0.9*
	(TFP, TVP)	(0.25, 0.5)	(0, 0)	(0, 0.75)

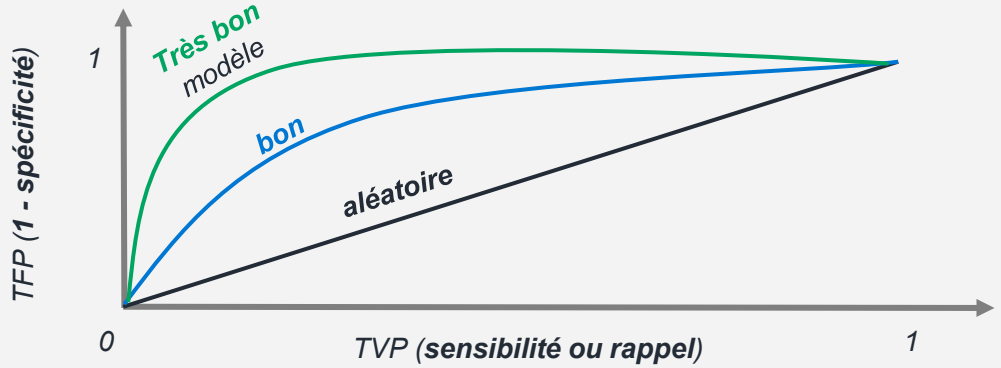
Exemple illustratif

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites

Approche 2 : AUC-ROC

L'**AUC-ROC** (Area Under the Curve - Receiver Operating Characteristic) ou **AUC** évalue la capacité d'un modèle à distinguer entre les classes en traçant le TVP contre le TFP **pour différents seuils** de décision. L'AUC (aire sous la courbe) varie entre 0,5 (modèle aléatoire) et 1 (modèle parfait)



Indice de GINI = 2 x AUC - 1

Des mesures **« agnostiques »** aux modèles (s'appliquent à toutes techniques : GLM, arbres de décision, …)

Evaluation empirique du modèle (1/3)

Coefficient de détermination linéaire de Pearson

- **Le coefficient de détermination R^2** mesure la proportion de la variance de la variable cible expliquée par les variables indépendantes dans un modèle linéaire. Il est utilisé pour évaluer la **qualité d'ajustement d'un modèle linéaire**. Il varie entre 0 (mauvais ajustement) et 1 (ajustement parfait)

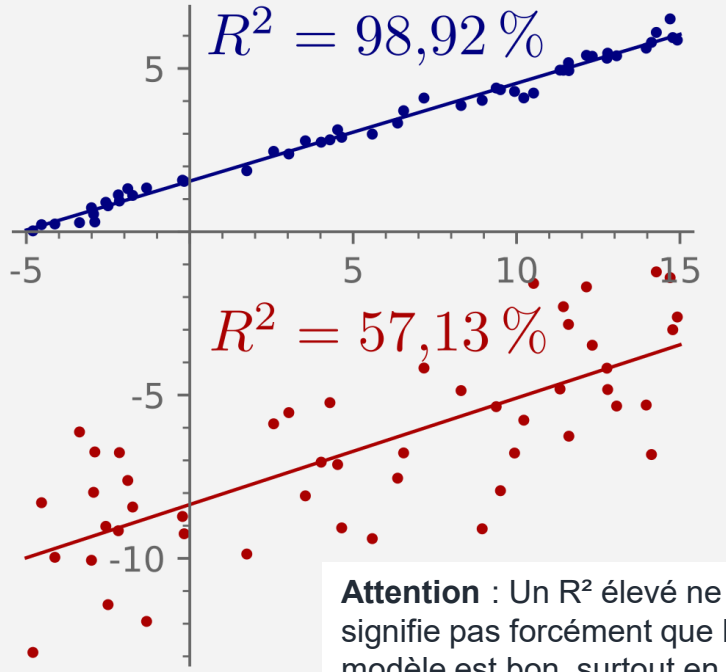
$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Avec Y_i valeur observée, $\hat{Y}_i$ valeur estimée et $\bar{Y}$ valeur moyenne

Interprétation :

- $R^2 = 1$: Le modèle explique **toute** la variance des données
- $R^2 = 0$: Le modèle n'explique **aucune** variance des données.
- $R^2 < 0$: Le modèle est **pire** qu'un modèle constant (moyenne).

$R^2 = 0.85$ signifie que 85 % de la variance de la variable cible est expliquée par le modèle



Attention : Un R² élevé ne signifie pas forcément que le modèle est bon, surtout en cas de données non linéaires.

Evaluation empirique du modèle (1/3)

Coefficient de détermination linéaire de Pearson

Le coefficient de détermination R^2 mesure la proportion de la variance de la variable cible expliquée par les variables indépendantes dans un modèle linéaire. Il est utilisé pour évaluer la **qualité d'ajustement d'un modèle linéaire**. Il varie entre 0 (mauvais ajustement) et 1 (ajustement parfait)

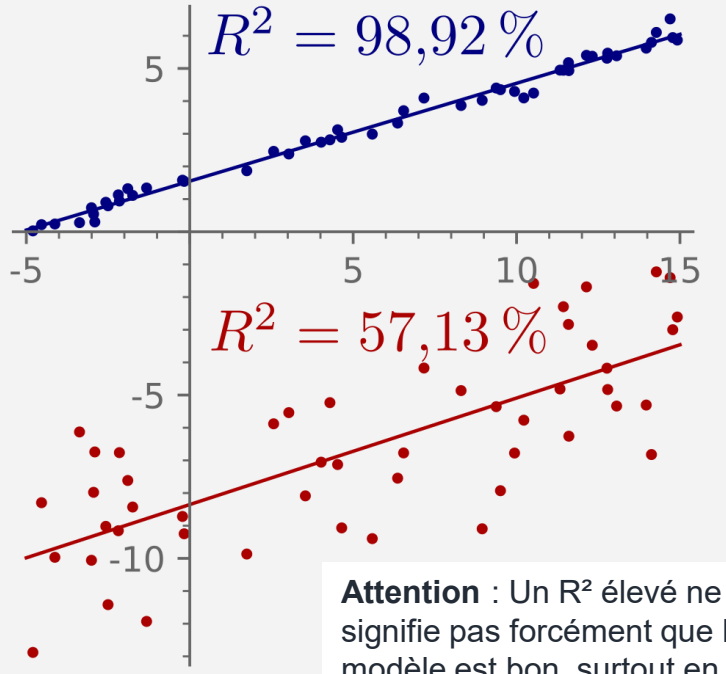
$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Avec Y_i valeur observée, $\hat{Y}_i$ valeur estimée et $\bar{Y}$ valeur moyenne

Interprétation :

- $R^2 = 1$: Le modèle explique **toute** la variance des données
- $R^2 = 0$: Le modèle n'explique **aucune** variance des données.
- $R^2 < 0$: Le modèle est **pire** qu'un modèle constant (moyenne).

$R^2 = 0.85$ signifie que 85 % de la variance de la variable cible est expliquée par le modèle



Attention : Un R² élevé ne signifie pas forcément que le modèle est bon, surtout en cas de données non linéaires.

Problème avec GLM : le R^2 des modèles linéaires est basé sur la somme des carrées des résidus et ne tient pas compte du **surajustement**. Il n'y a donc pas de R^2 dans les GLM qui sont basés sur le **Maximum de Vraisemblance.**

Les *pseudos-R^2* sont des **équivalents du R^2 dans les GLM**

$$\textit{Pseudo-}R^2 = 1 - \frac{\textit{déviance modèle}}{\textit{déviance nulle}} = 1 - \frac{(-2\log \textit{vraisemblance modèle})}{(-2\log \textit{vraisemblance nulle})}$$

ISFA Institut de Science Financière et d'Assurances

Evaluation empirique du modèle (2/3)

Les pseudos R^2 : quelques indicateurs

Comparer le modèle avec le modèle initial (trivial) constitué de la seule constante. Soit $L_{modèle}$ la **vraisemblance maximale du modèle ajusté** (avec les variables explicatives) et L_0 la **vraisemblance maximale du modèle nul** (sans variables explicatives, uniquement avec l'intercept).

R^2 de Cox et Snell

mesure pseudo-R² adaptée aux modèles de régression non linéaire, comme la régression logistique. Il est basé sur le rapport de vraisemblance et exprime la proportion de la variance expliquée par le modèle. Formulé comme :

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_{modèle}}\right)^{\frac{2}{n}}$$

Ce R^2 est borné par 1, mais ne l'atteint jamais exactement, ce qui peut limiter son interprétation en comparaison avec d'autres mesures comme le **R^2 de Nagelkerke**.

$$R^2_N = \frac{R^2_{CS}}{\max(R^2_{CS})}$$

R^2 de McFadden

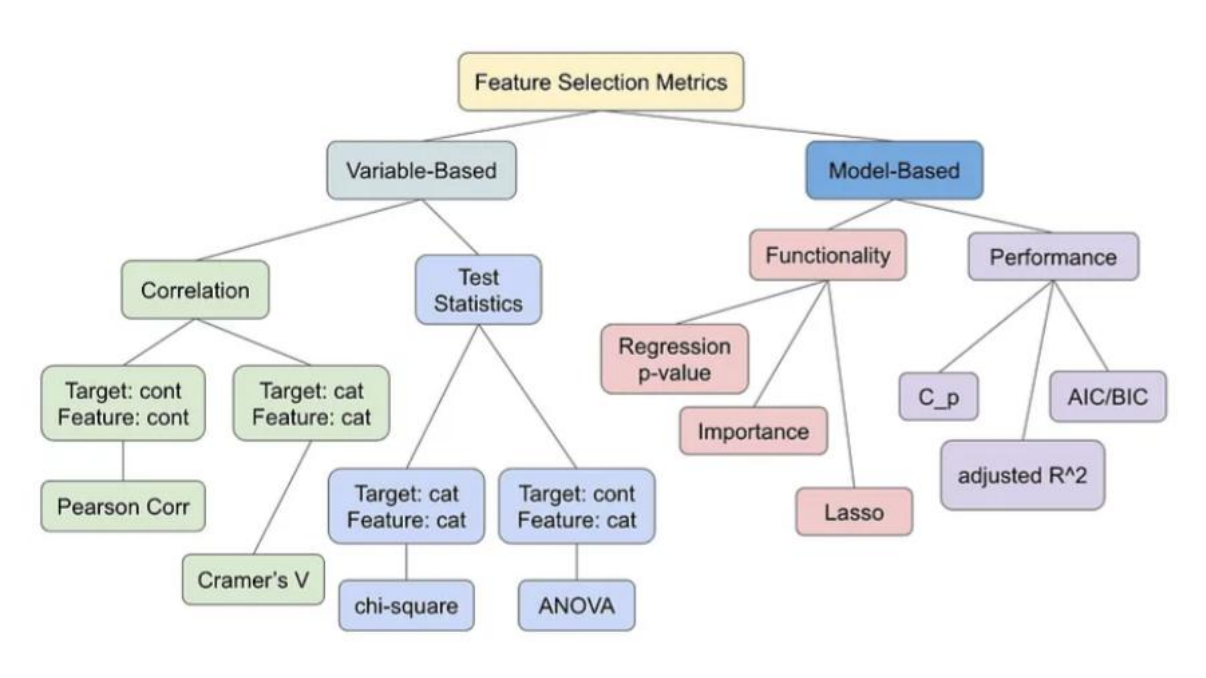
Evalue la qualité d'ajustement d'un modèle à partir de la log-vraisemblance.

$$R^2_{MF} = 1 - \frac{\log(L_{modèle})}{\log(L_0)}$$

Par rapport au R^2 de Cox et Snell ou à celui de Nagelkerke, le R^2 de McFadden est **plus couramment utilisé pour les modèles logistiques** car il repose sur des principes de vraisemblance qui s'ajustent bien à ces modèles.

Souvent préféré pour la régression logistique en raison de son **lien direct avec le test du rapport de vraisemblance**.

7. Sélection des variables

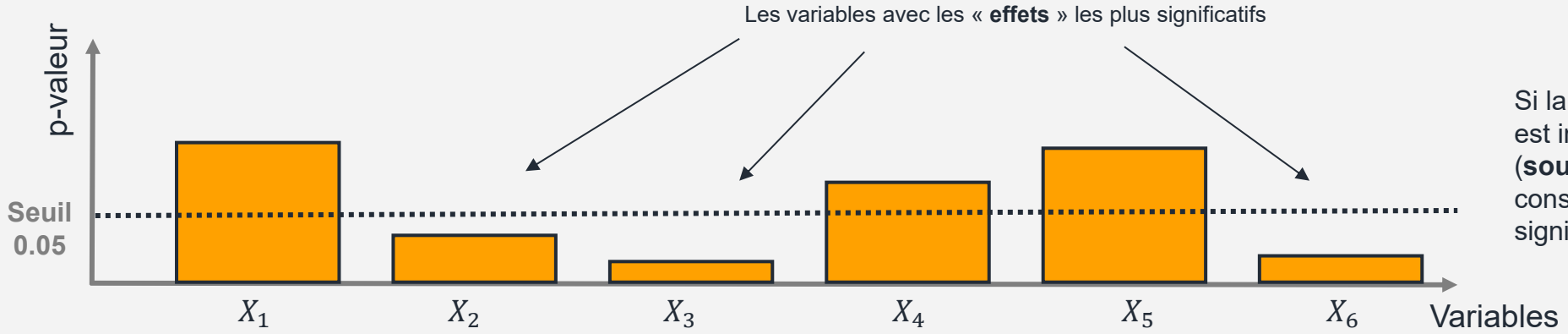


ISFA
Institut de Science Financière et d'Assurances

Tests statistiques (1/2)

Test du rapport de vraisemblance

Exemple illustratif



Si la **p-valeur** associée au coefficient est inférieure à un seuil conventionnel (**souvent 0.05**), la variable est considérée comme ayant un effet significatif sur la variable cible.

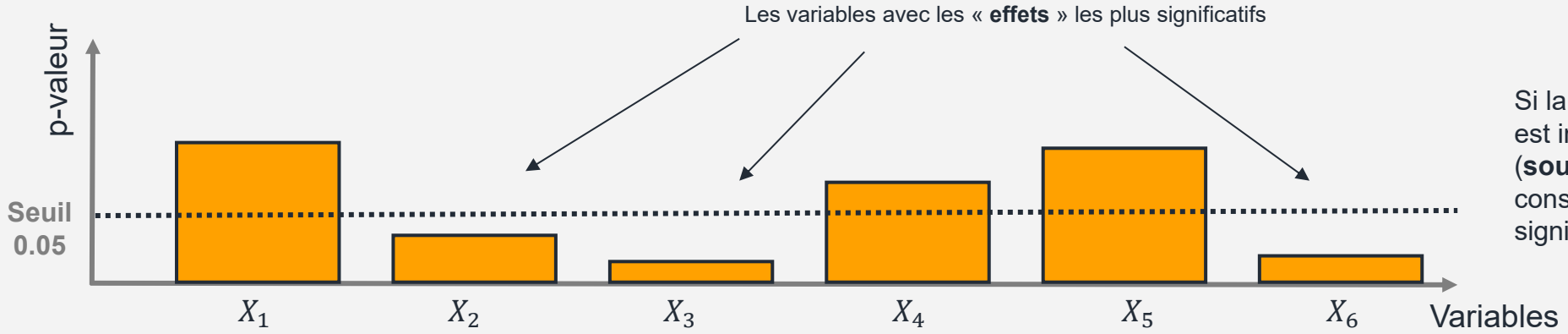
- Test statistique : $\boldsymbol{\beta_j = 0}$ vs $\boldsymbol{\beta_j \neq 0}$ avec $j \in \{1, \dots, d\}$

Le prédicteur n'est **pas significatif** dans le GLM

Tests statistiques (1/2)

Test du rapport de vraisemblance

Exemple illustratif



Si la **p-valeur** associée au coefficient est inférieure à un seuil conventionnel (**souvent 0.05**), la variable est considérée comme ayant un effet significatif sur la variable cible.

Test statistique : $\boldsymbol{\beta_j = 0}$ vs $\boldsymbol{\beta_j \neq 0}$ avec $j \in \{1, \dots, d\}$

▶ Rapport de vraisemblance

$$\mathrm{LR_j} = D_{-j} - D_{modèle} \sim \boldsymbol{\chi^2(1)} \text{ asymptotiquement}$$

Avec la **déviance du modèle ajusté** (avec toutes les variables explicatives)

$$D_{Modèle} = -2 \times log(L_{Modèle})$$

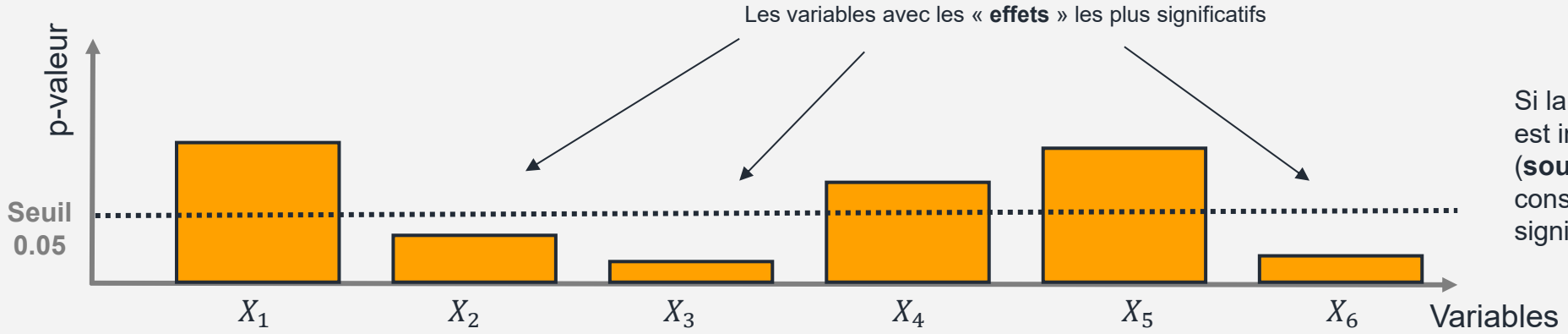
la **déviance du modèle réduit** (toutes les variables sauf X_j)

$$D_{-j} = -2 \times log(L_{-j})$$

Tests statistiques (1/2)

Test du rapport de vraisemblance

Exemple illustratif



Si la **p-valeur** associée au coefficient est inférieure à un seuil conventionnel (**souvent 0.05**), la variable est considérée comme ayant un effet significatif sur la variable cible.

Test statistique : $\boldsymbol{\beta_j = 0}$ vs $\boldsymbol{\beta_j \neq 0}$ avec $j \in \{1, \dots, d\}$

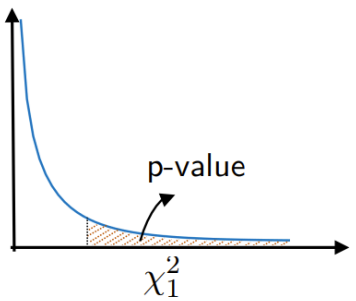
Rapport de vraisemblance

$$\mathrm{LR_j} = D_{-j} - D_{modèle} \sim \boldsymbol{\chi^2(1)} \text{ asymptotiquement}$$

- p-valeur : $1 - F_{\chi^2(1)}(LR_j)$

Avec **la fonction de répartition** :

$$F_{\chi^2(1)}(u) = \mathbb{P}(\chi^2(1) \leq u)$$

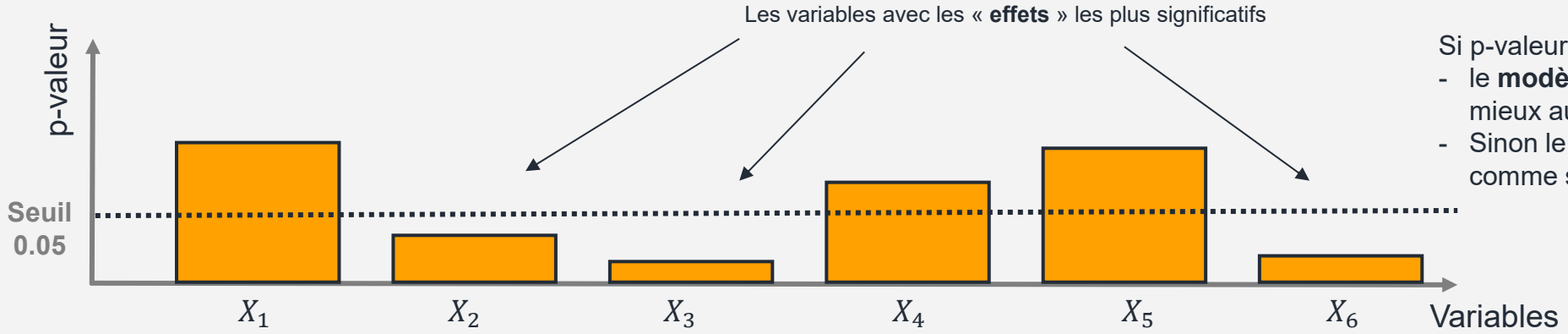


ISFA
Institut de Science Financière et d'Assurances

Tests statistiques (1/2)

Test du rapport de vraisemblance, évaluer un groupe de prédicteurs

Exemple illustratif



Si p-valeur < 0.05 alors:
- le **modèle complet** s'ajuste significativement mieux aux données que le modèle restreint.
- Sinon le **modèle restreint** est considéré comme suffisant pour expliquer les données.

Test statistique : $\boldsymbol{\beta_j = 0}$ vs $\boldsymbol{\beta_j \neq 0}$ avec $j \in \{1, \dots, d\}$

Rapport de vraisemblance

$\mathrm{LR_j} = D_{-j} - D_{modèle} \sim \boldsymbol{\chi^2(1)}$ asymptotiquement

p-valeur : $1 - F_{\chi^2(1)}(LR_j)$

▶ Test statistique : $\boldsymbol{\beta_1, \beta_2, \dots, \beta_r = 0}$ vs $\boldsymbol{\beta_1, \beta_2, \dots, \beta_r \neq 0}$ avec $\mathrm{r} \leq d$

Rapport de vraisemblance

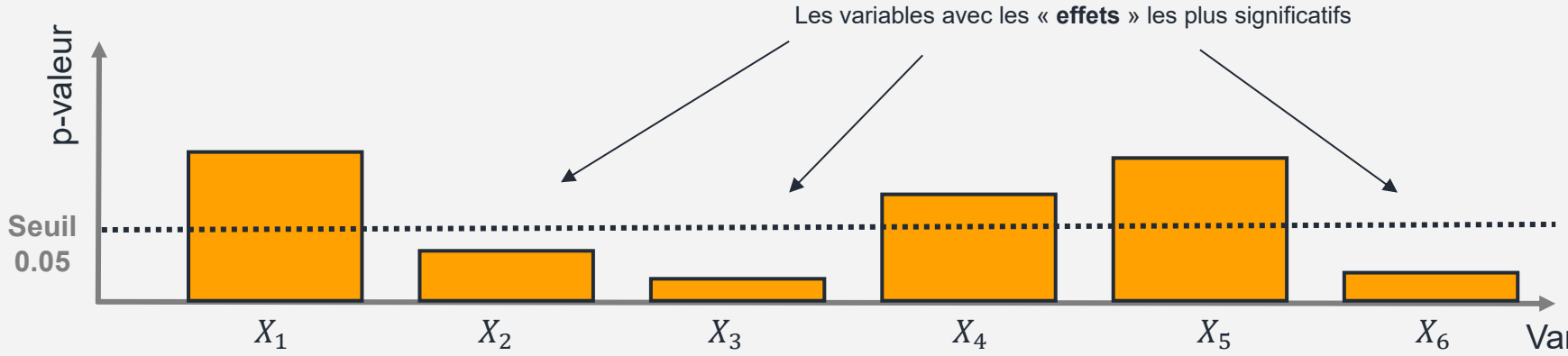
$\mathrm{LR} = D_{-(1,\dots,r)} - D_{Modèle} \sim \boldsymbol{\chi^2(d-r)}$ asymptotiquement

p-valeur : $1 - F_{\chi^2(d-r)}(LR)$

Tests statistiques (2/2)

Test de Wald : moins précis mais moins coûteux en ressources

Exemple illustratif



Si la **p-valeur** associée au coefficient est inférieure à un seuil conventionnel (**souvent 0.05**), la variable est considérée comme ayant un effet significatif sur la variable cible.

Test statistique : $\boldsymbol{\beta_j = 0}$ vs $\boldsymbol{\beta_j \neq 0}$ avec $j \in \{1, \dots, d\}$

Approche moins précise mais en même temps moins coûteuse en ressources

Rapport de vraisemblance

$$\mathrm{LR_j} = D_{-j} - D_{modèle} \sim \boldsymbol{\chi^2(1)} \text{ asymptotiquement}$$

p-valeur : $1 - F_{\chi^2(1)}(LR_j)$

▶ **Test de Wald**, avec $\widehat{\sigma_j}$ l'erreur standard de $\widehat{\beta_j}$

$$\mathrm{W_j} = \frac{\widehat{\beta_j}^2}{Var(\widehat{\beta_j})} \sim \boldsymbol{\chi^2(1)} \text{ asymptotiquement}$$

p-valeur : $1 - F_{\chi^2(1)}(W_j)$

Critères AIC et BIC

Des critères de sélection qui tiennent en compte le nombre de variables explicatives

Exemple illustratif

Constat : L'ajout de variables dans un modèle **réduit la déviance (ou augmente la vraisemblance)**,

$$D_{modèle} = -2 \times log(L_{modèle})$$

même si certaines variables sont non pertinentes, similaire à l'augmentation du R^2 en régression linéaire avec une diminution des degrés de liberté.

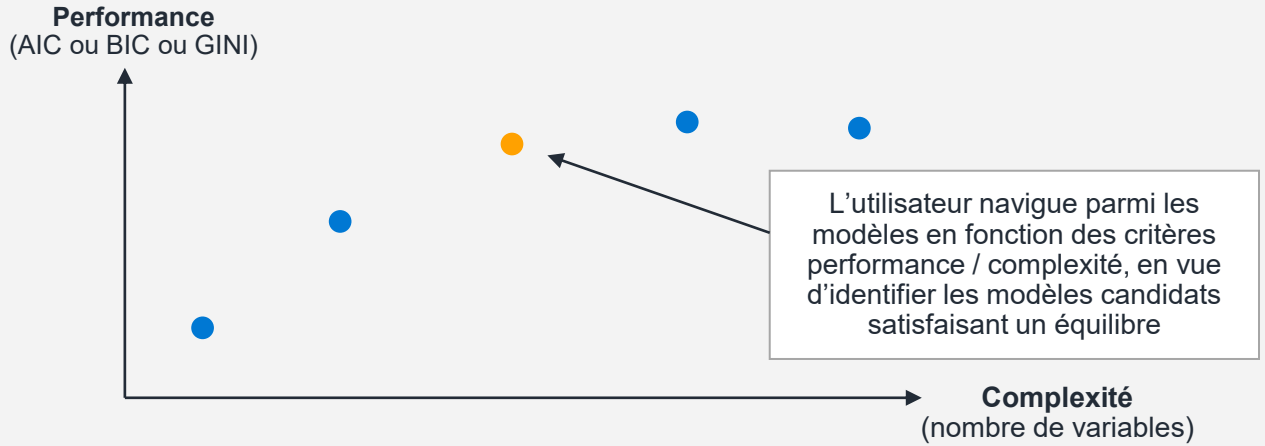
Solution : Contrebalancer la réduction de la déviance par une **pénalisation liée à la complexité du modèle**, comme avec l'AIC ou le BIC.

Critère AIKAKE :

$$AIC = -2\log(L_{modèle}) + 2 \times (d+1)$$

Critère Schwartz :

$$BIC = -2\log(L_{modèle}) + \log(n) \times (d+1)$$



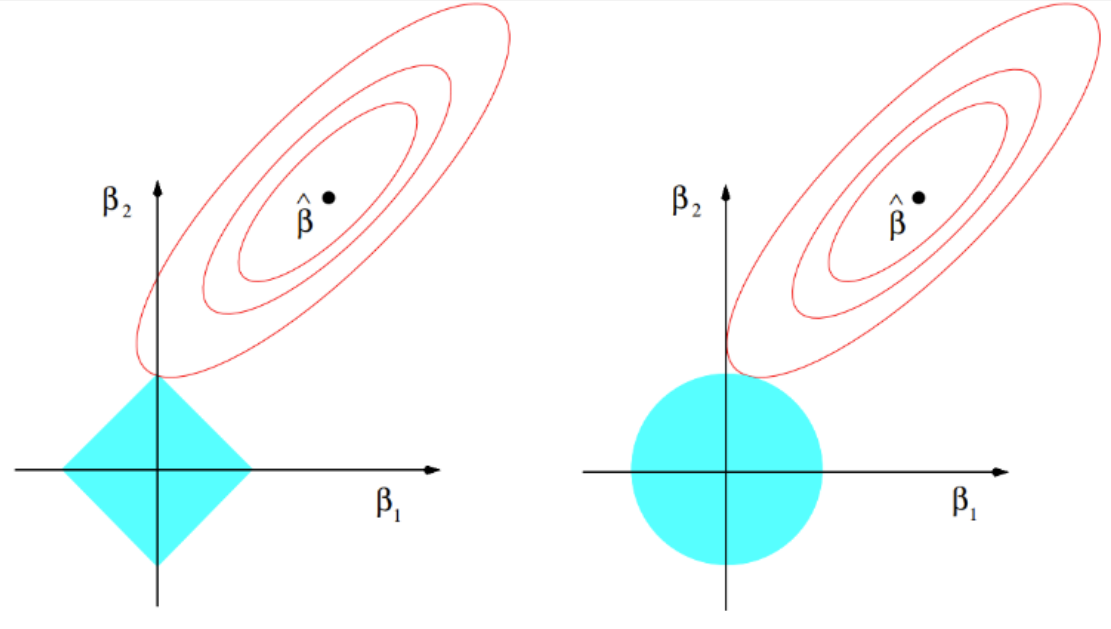
PROCEDURE

Nous allons évaluer des successions de modèles emboîtés :

- **FORWARD :** en les ajoutant au fur et à mesure
- **BACKWARD :** en les retirant au fur et à mesure
- **STEPWISE :** en alternant FORWARD / BACKWARD

Règle d'arrêt : l'ajout ou le retrait d'une variable n'améliore plus le critère

8. GLM pénalisé



ISFA
Institut de Science Financière et d'Assurances

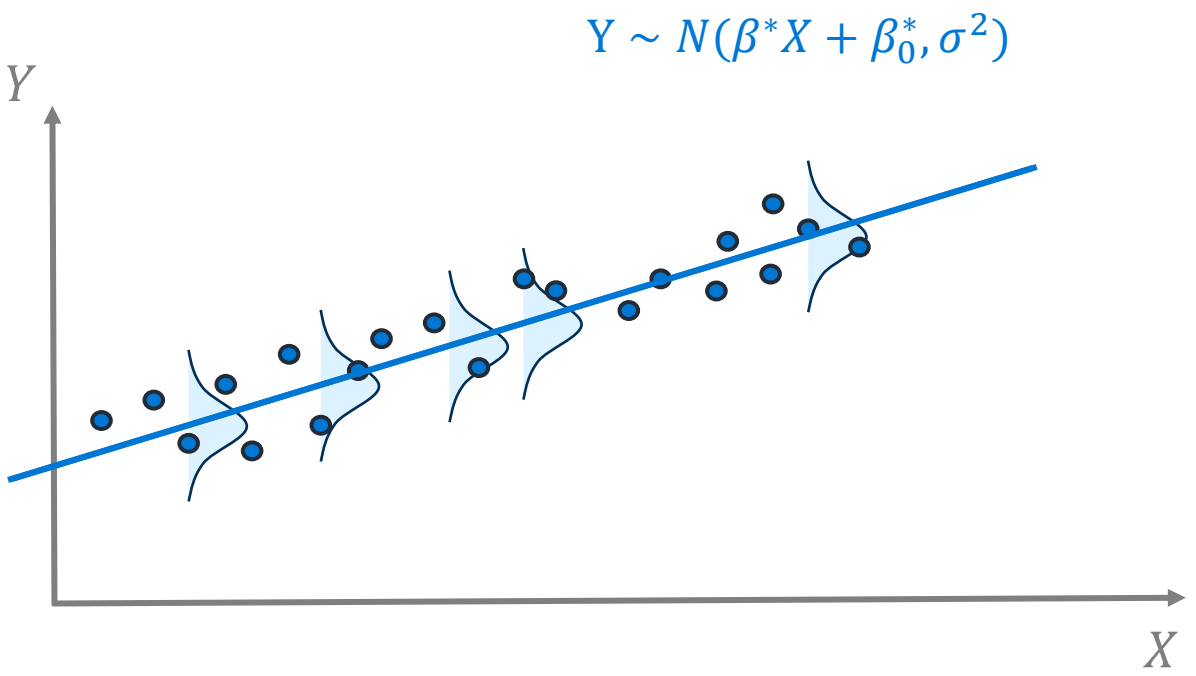
Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) :
chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

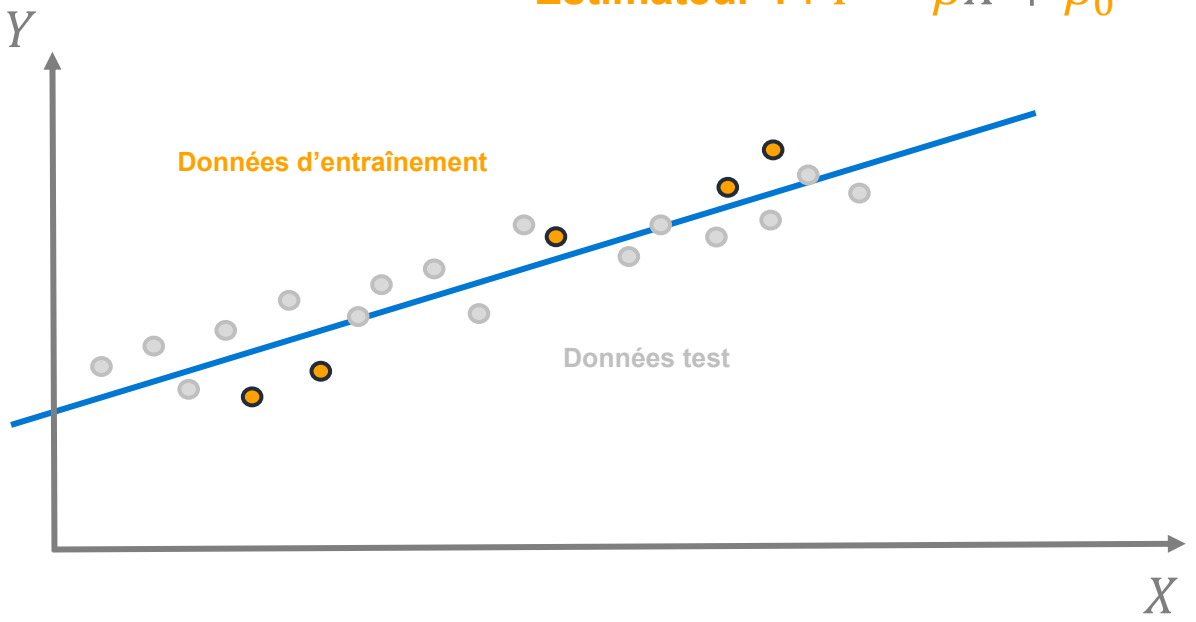
Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$

$$Y \sim N(\beta^* X + \beta_0^*, \sigma^2)$$

Estimateur 1 : $\hat{Y} = \beta X + \beta_0$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

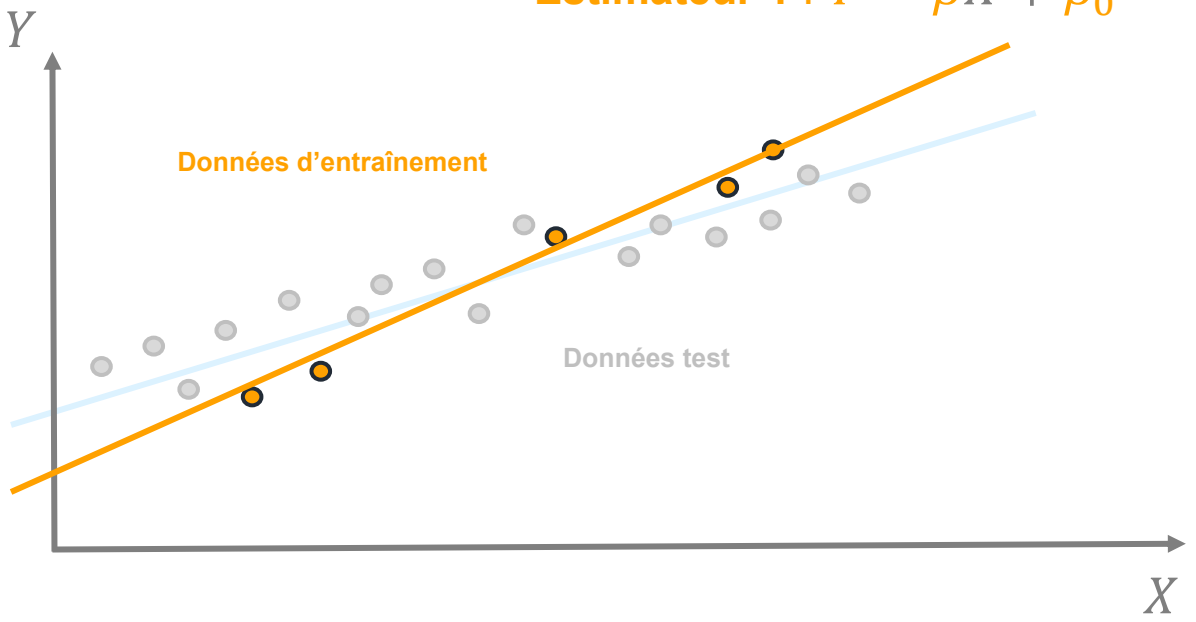
Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$

$$Y \sim N(\beta^* X + \beta_0^*, \sigma^2)$$

Estimateur 1 : $\hat{Y} = \beta X + \beta_0$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

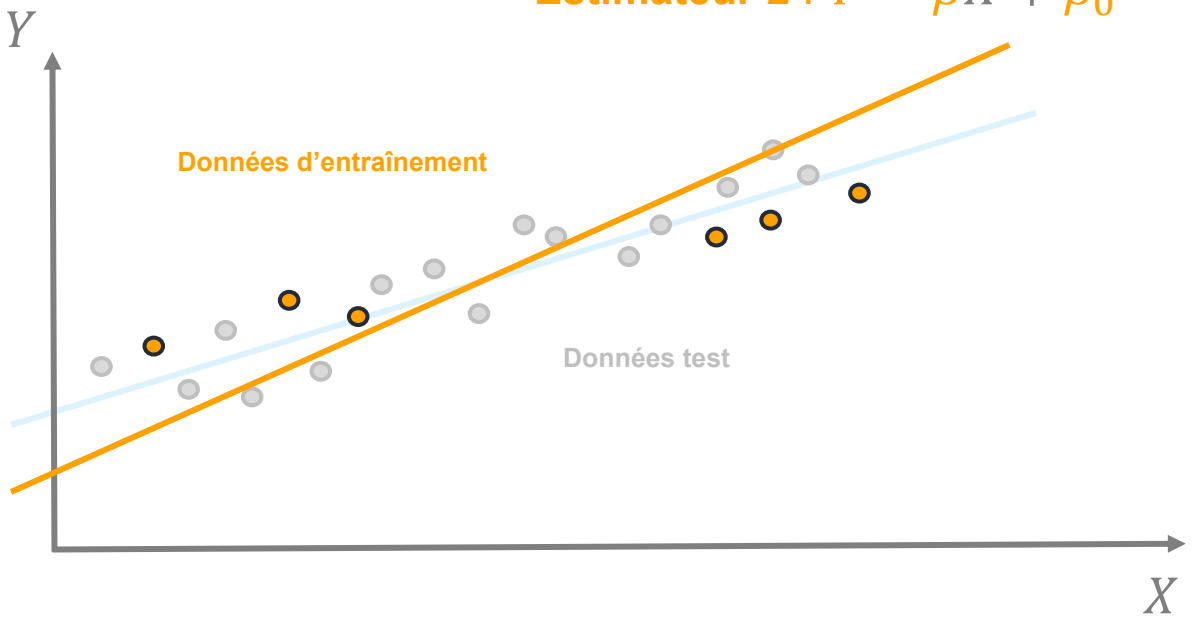
Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$

$$Y \sim N(\beta^* X + \beta_0^*, \sigma^2)$$

Estimateur 2 : $\hat{Y} = \beta X + \beta_0$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

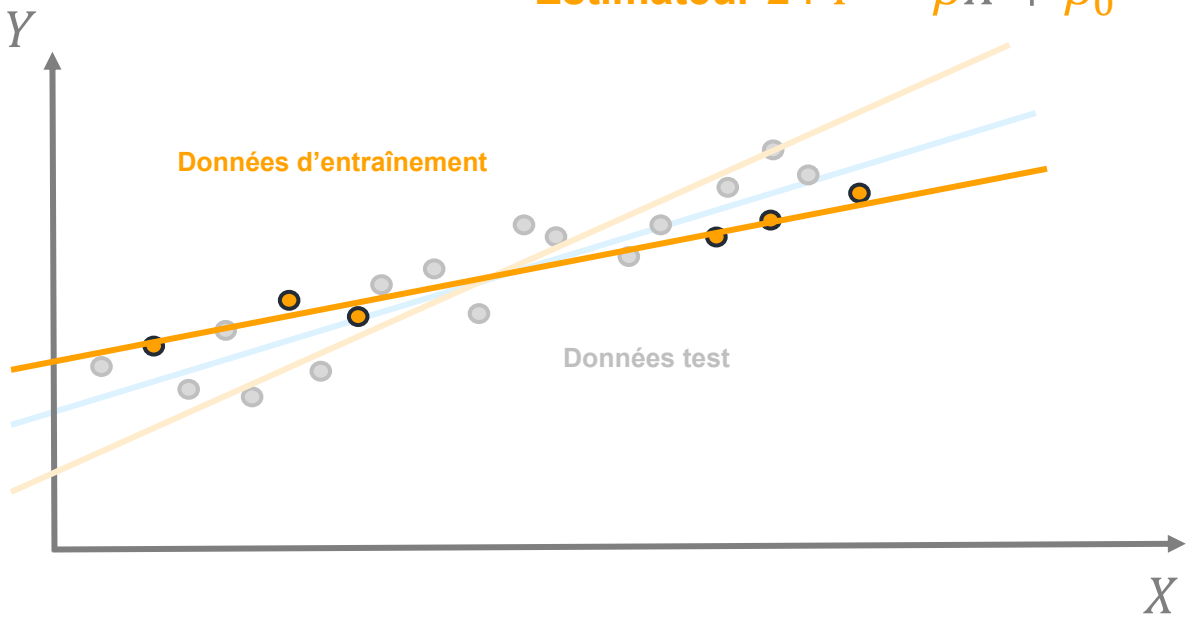
Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$

$$Y \sim N(\beta^* X + \beta_0^*, \sigma^2)$$

Estimateur 2 : $\hat{Y} = \beta X + \beta_0$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta_i^T}\right)^2$$

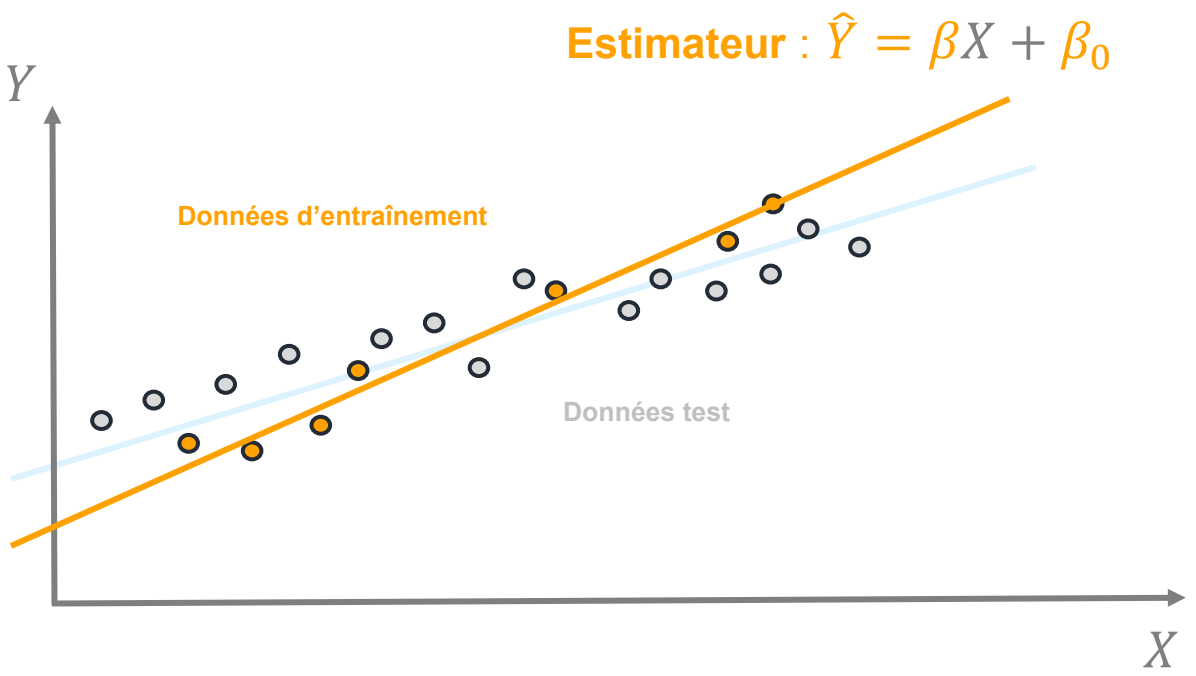
- **Modèle linéaire pénalisé (Ridge ou Lasso)** : modèle linéaire qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta_{ridge}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} \boldsymbol{\beta_i^2}$$

$$\boldsymbol{\beta_{lasso}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} |\, \boldsymbol{\beta_i} \,|$$



Problème : Incapacité à gérer le sur-apprentissage (ou surajustement)

Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta}_i^T\right)^2$$

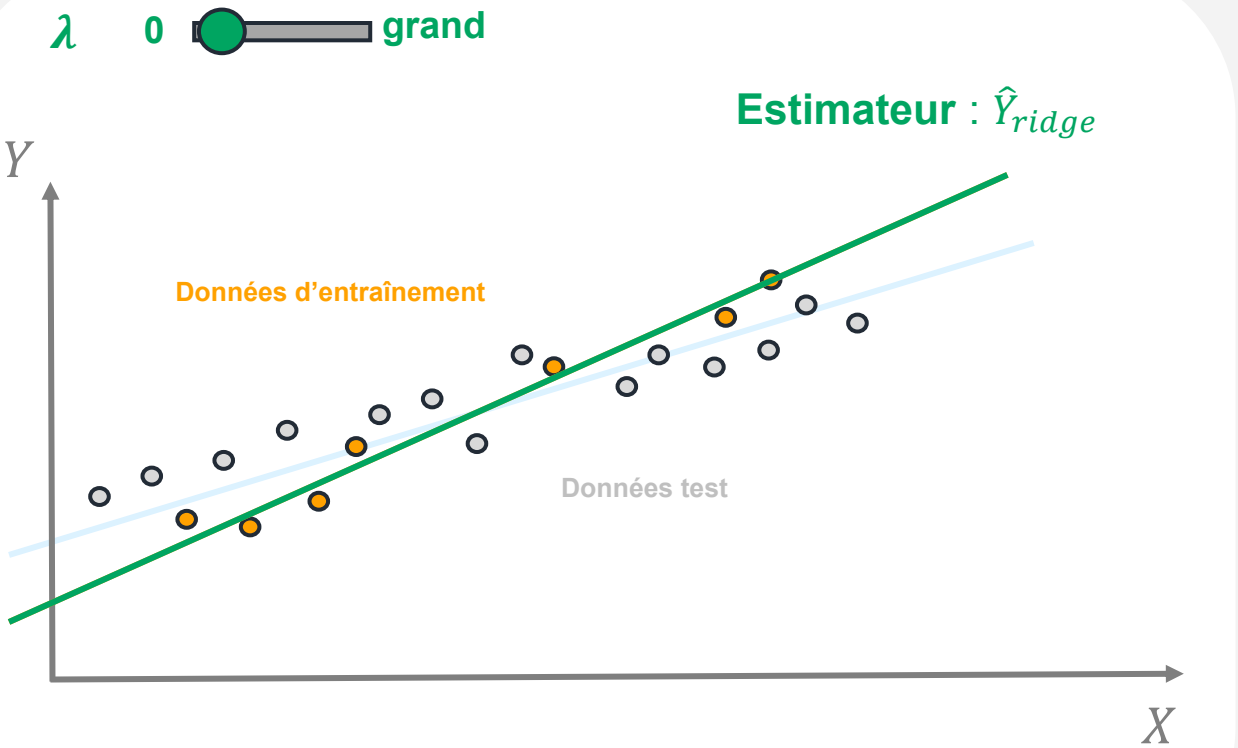
- **Modèle linéaire pénalisé (Ridge ou Lasso)** : modèle linéaire qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta_{ridge}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} \boldsymbol{\beta}_i^2$$

$$\boldsymbol{\beta_{lasso}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} |\, \boldsymbol{\beta_i} \,|$$



Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta}_i^T\right)^2$$

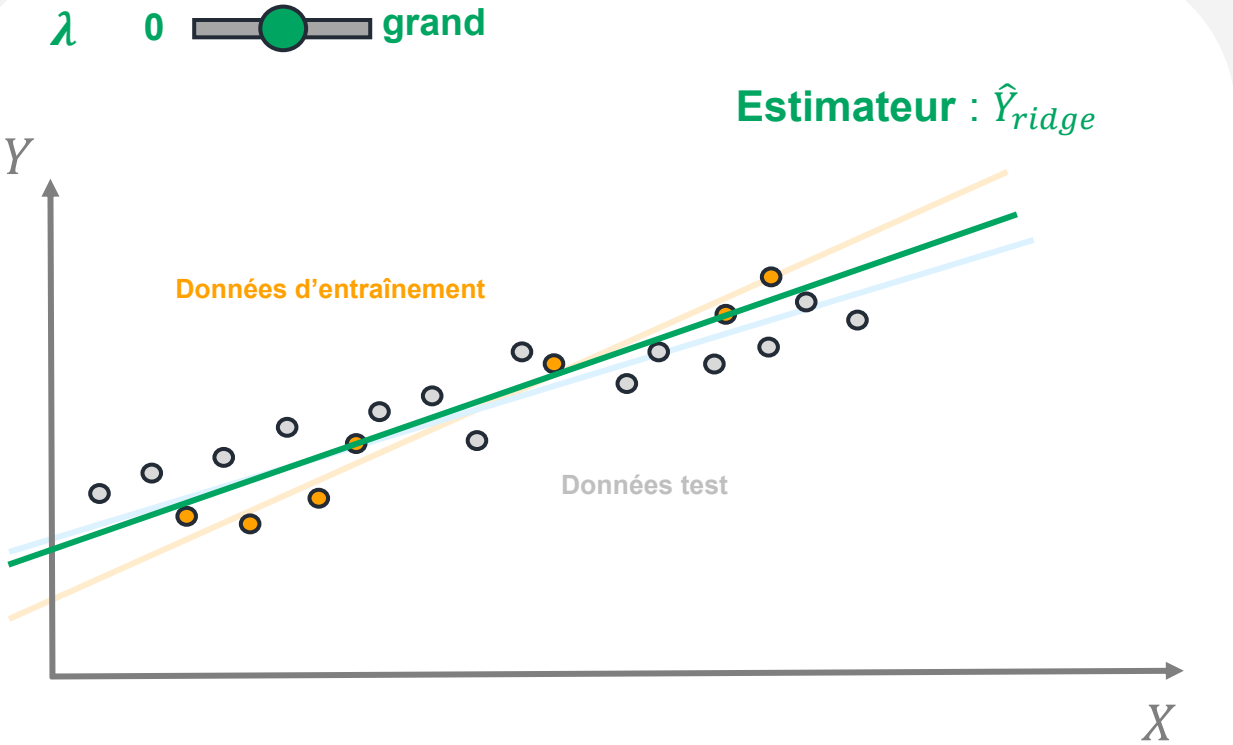
- **Modèle linéaire pénalisé (Ridge ou Lasso)** : modèle linéaire qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta_{ridge}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} \boldsymbol{\beta}_i^2$$

$$\boldsymbol{\beta_{lasso}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} |\, \boldsymbol{\beta_i} \,|$$



Formulation des GLM pénalisés (1/2)

Cas des modèles linéaires classiques

Exemple illustratif

Rappel : Estimateur **des moindres carrés** (*OLS*) : chercher les valeurs des paramètres minimisant

$$\boldsymbol{\beta_{OLS}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \left(Y_i - X_i \boldsymbol{\beta}_i^T\right)^2$$

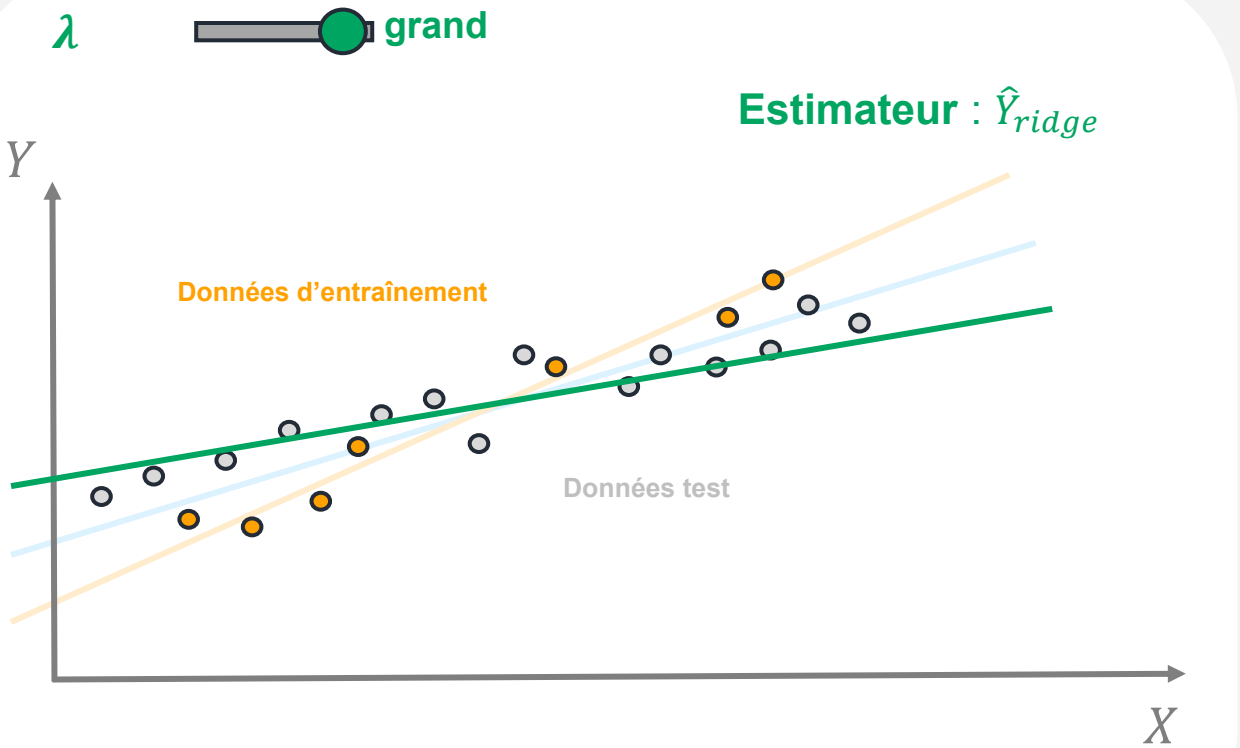
Modèle linéaire pénalisé (Ridge ou Lasso) : modèle linéaire qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta_{ridge}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} \boldsymbol{\beta}_i^2$$

$$\boldsymbol{\beta_{lasso}} = \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

$$= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + \boldsymbol{\lambda} \sum_{i=1,\dots,n} |\, \boldsymbol{\beta_i} \,|$$



▶ Nous sélectionnons la pénalisation optimale avec la **validation croisée !**

Formulation des GLM pénalisés (2/2)

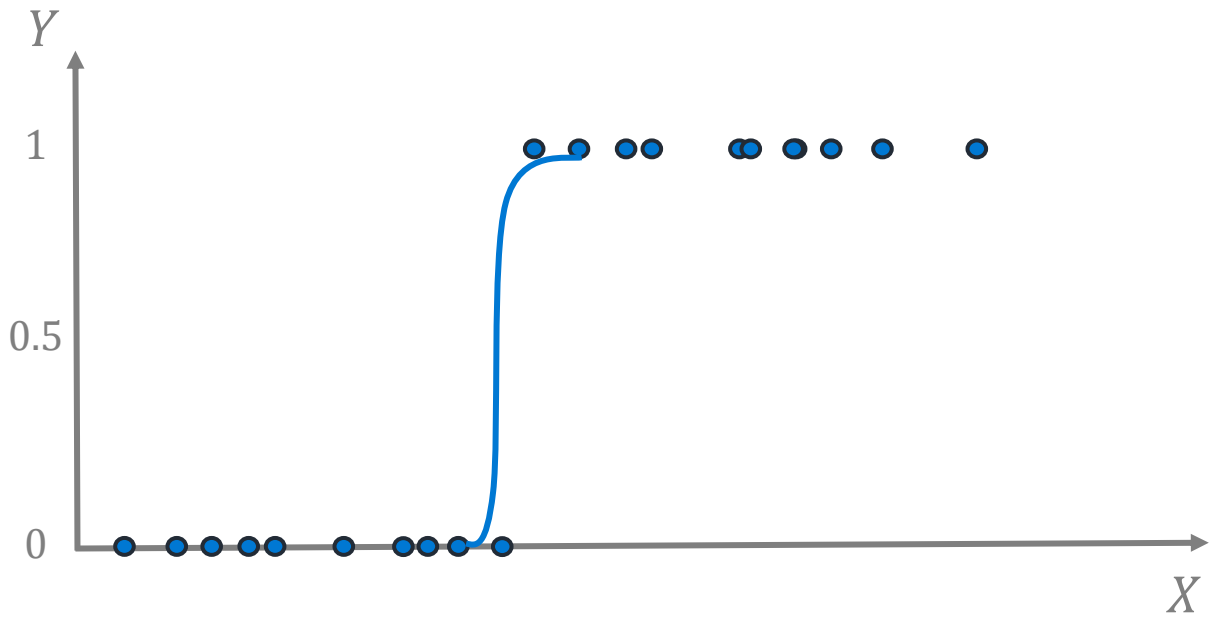
Cas des modèles linéaires généralisés

Exemple illustratif

Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

$$Y \sim Ber(p^*)$$



Formulation des GLM pénalisés (2/2)

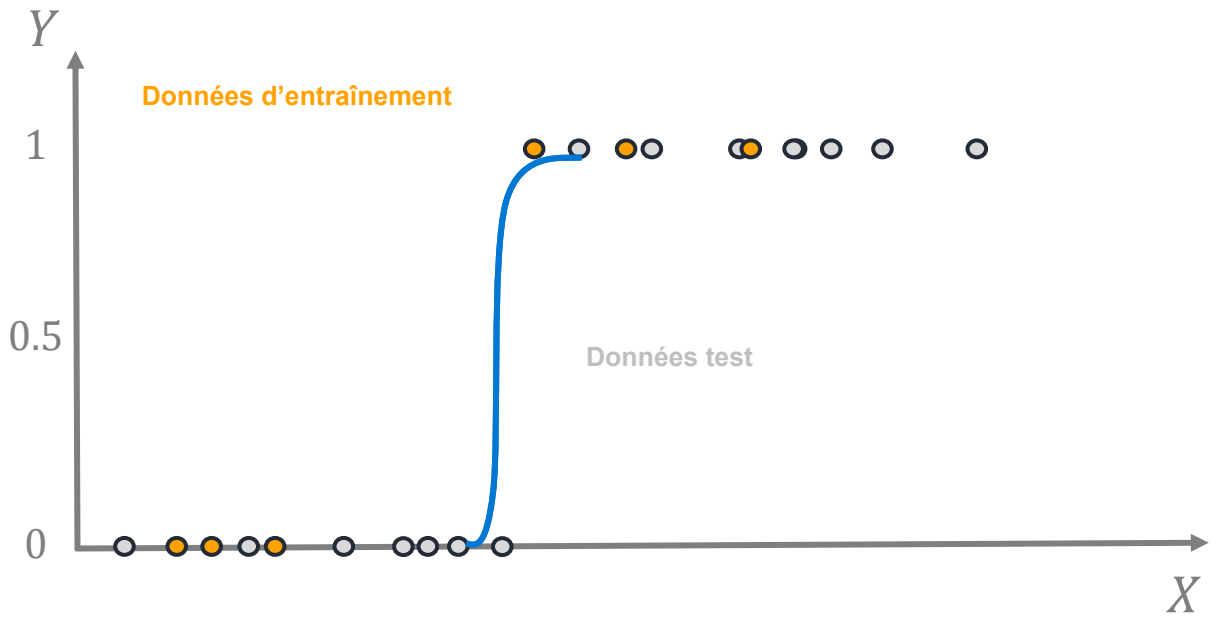
Cas des modèles linéaires généralisés

Exemple illustratif

Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

$Y \sim Ber(p^*)$



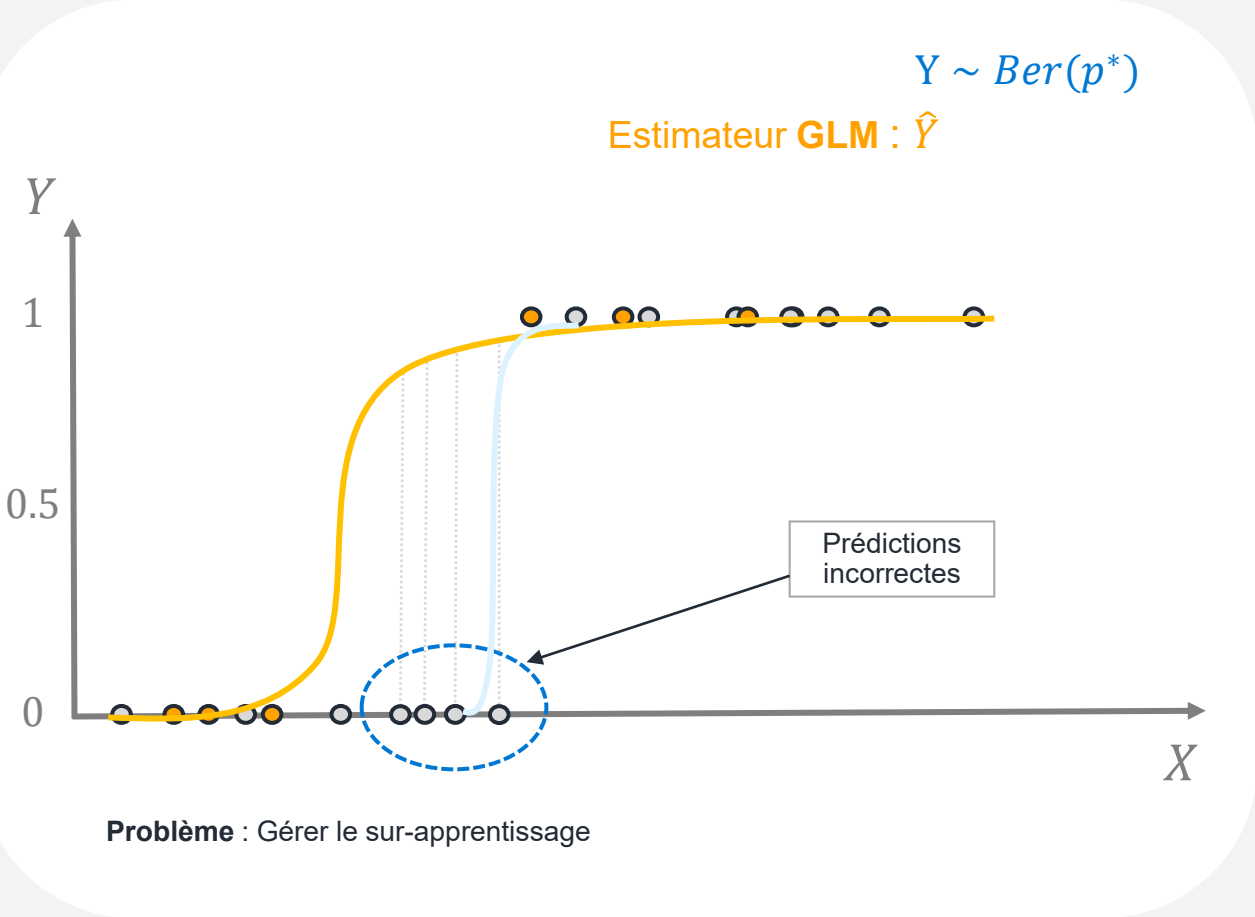
Formulation des GLM pénalisés (2/2)

Cas des modèles linéaires généralisés

Exemple illustratif

Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$



Problème : Gérer le sur-apprentissage

Formulation des GLM pénalisés (2/2)

Cas des modèles linéaires généralisés

Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

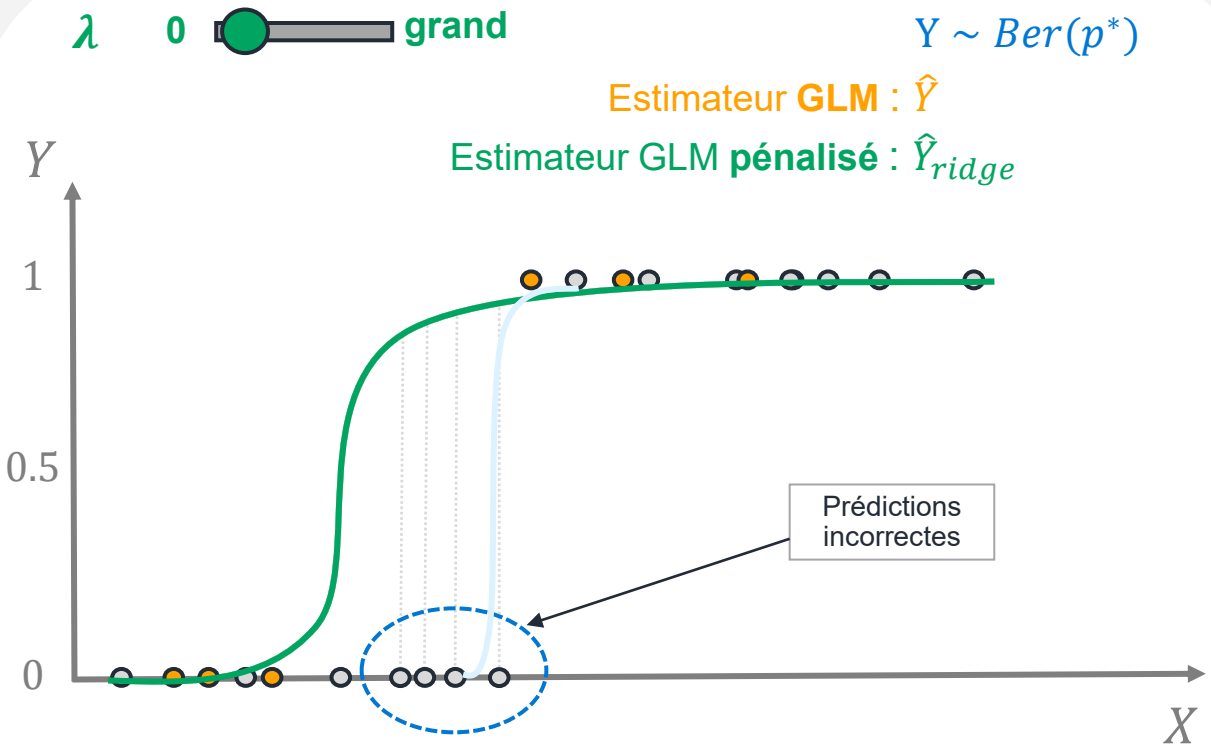
$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

- **GLM avec une pénalité (Ridge ou Lasso)** : GLM qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta}_{MLE-ridge} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$\boldsymbol{\beta}_{MLE-lasso} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

Exemple illustratif



Problème : Gérer le sur-apprentissage

Formulation des GLM pénalisés (2/2)

Cas des modèles linéaires généralisés

Exemple illustratif

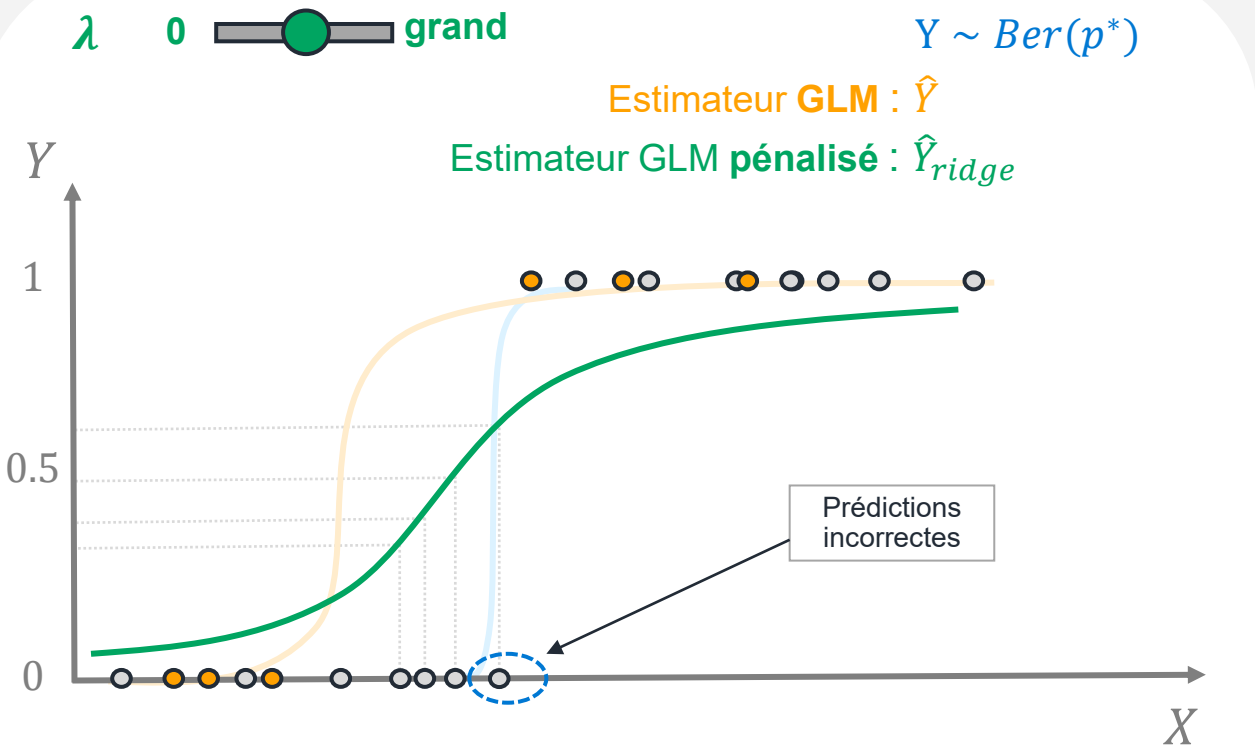
Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

- **GLM avec une pénalité (Ridge ou Lasso)** : GLM qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta}_{MLE-ridge} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

$$\boldsymbol{\beta}_{MLE-lasso} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$



Formulation des GLM pénalisés (2/2)

Cas des modèles linéaires généralisés

Exemple illustratif

Rappel : Estimateur du **maximum de vraisemblance** (*MLE*) : chercher les valeurs des paramètres maximisant **la fonction de vraisemblance** (ou **log de vraisemblance**)

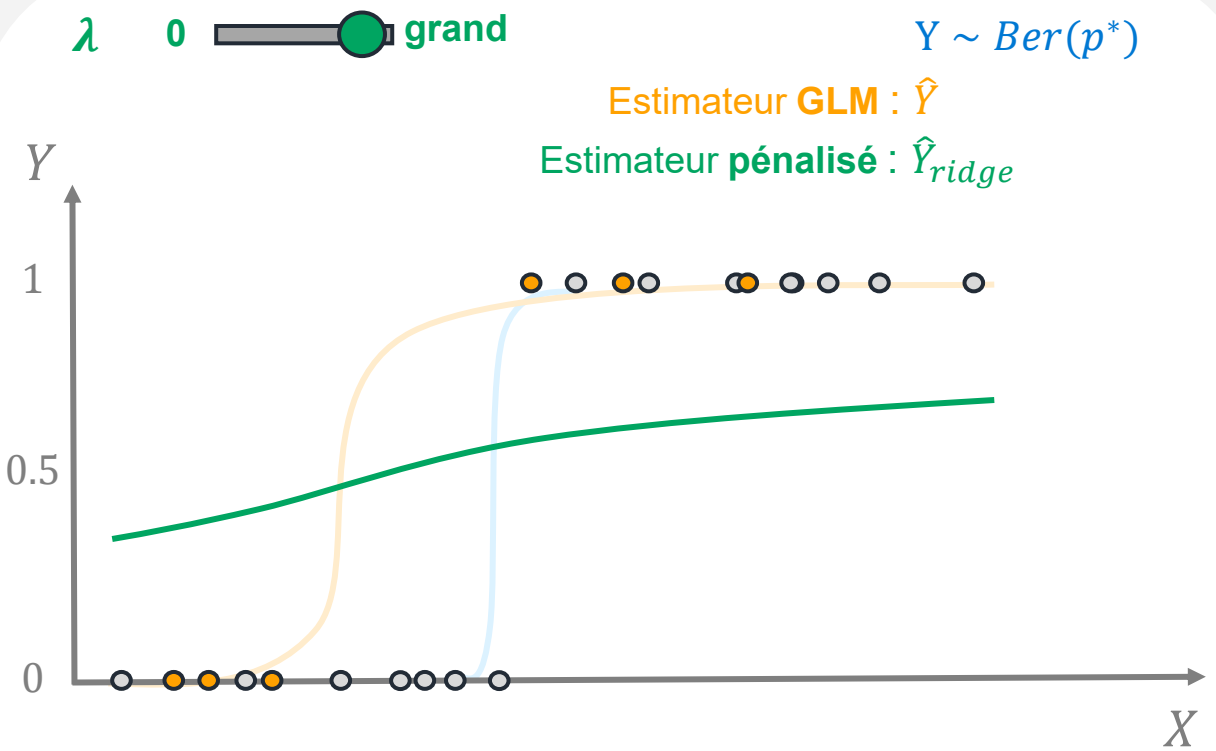
$$\boldsymbol{\beta}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1,\dots,n} \log f_{\boldsymbol{\beta}}(y_i)$$

GLM avec une pénalité (Ridge ou Lasso) : GLM qui ajoute **une pénalité sur la magnitude des coefficients** pour éviter le surajustement.

$$\boldsymbol{\beta}_{MLE-ridge} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^2$$

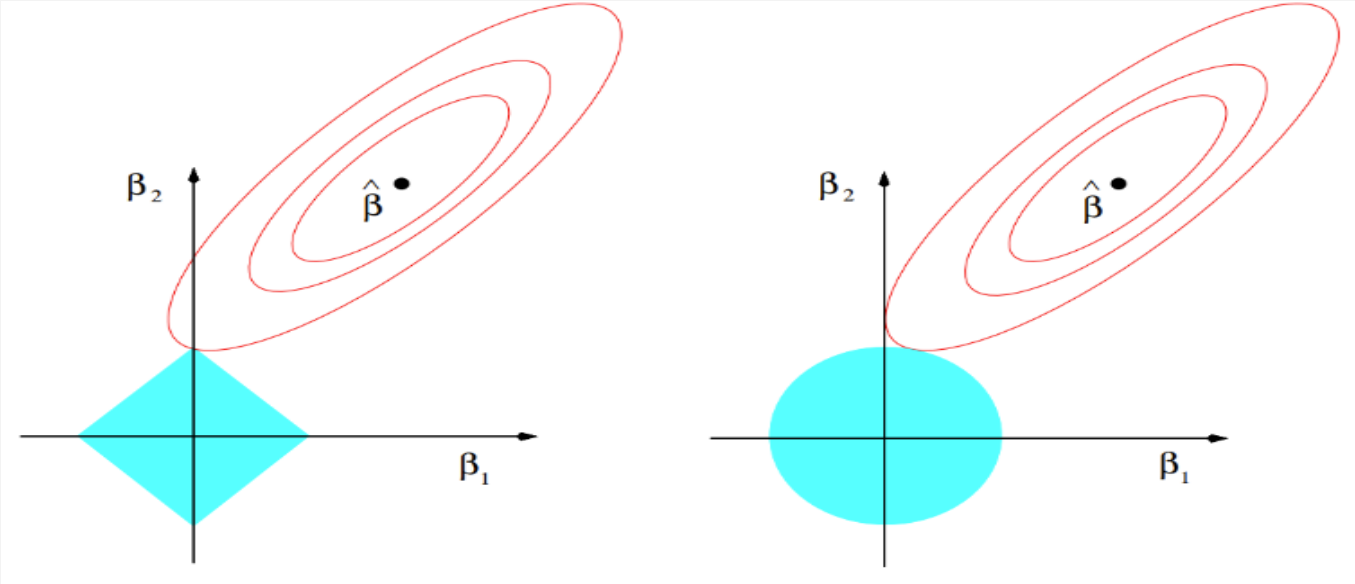
$$\boldsymbol{\beta}_{MLE-lasso} = \arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \boldsymbol{\lambda} \,||\, \boldsymbol{\beta} \,||^1$$

- Nous sélectionnons la pénalisation optimale avec la **validation croisée !**



Différences entre les modèles Ridge et Lasso (1/3)

Eviter le sur-apprentissage durant l'apprentissage



Lasso : La pénalité Lasso utilise la norme L1, **poussant certains coefficients à zéro**, ce qui permet une **sélection automatique des variables.** Il est utile pour simplifier les modèles en réduisant la dimensionnalité et en améliorant la généralisation.

VS

Ridge : La pénalité Ridge utilise la norme L2 pour **réduire la taille des coefficients**, aidant à améliorer la généralisation, surtout **en présence de variables corrélées**.

Différences entre les modèles Ridge et Lasso (2/3)

Tableau récapitulatif

	Ridge	Lasso
Type pénalité	**Norme L2** (somme des carrés des coefficients)	**Norme L1** (somme des valeurs absolues des coefficients)
Effet sur les coefficients	**Réduit les coefficients** mais ne les annule pas	Peut amener **certains coefficients à zéro**
Sélection de variables	✗	✓
Traiter la multi-colinéarité	✓	**Moins efficace** que Ridge
Stabilité des coefficients	✓	**Moins stable** que Ridge, (des coefficients sont nuls)
Utilisation optimale	Idéal lorsque les **variables sont corrélées** et toutes utiles	Idéal lorsqu'il y a **beaucoup de variables**, dont certaines non pertinentes
Quand utiliser ?	**Réduire la variance** sans éliminer de variables	**Simplifier le modèle** et à réduire la dimensionnalité

Différences entre les modèles Ridge et Lasso (3/3)

Extension avec la méthode Elastic-Net

- **Elastic Net** combine les pénalités **Ridge** (L2) et **Lasso** (L1) :

$$\arg\max_{\boldsymbol{\beta}} l(y, \boldsymbol{\beta}) - \lambda_2 \,||\, \boldsymbol{\beta} \,||^2 - \lambda_1 \,||\, \boldsymbol{\beta} \,||^1$$

- **Régularise les coefficients** tout en effectuant une **sélection de variables**
- Particulièrement utile lorsque les variables explicatives sont corrélées et qu'on souhaite à la fois réduire la dimensionnalité et éviter la multi-colinéarité.

	Ridge	Lasso
Type pénalité	**Norme L2** (somme des carrés des coefficients)	**Norme L1** (somme des valeurs absolues des coefficients)
Effet sur les coefficients	**Réduit les coefficients** mais ne les annule pas	Peut amener **certains coefficients à zéro**
Sélection de variables	✗	✓
Traiter la multi-colinéarité	✓	**Moins efficace** que Ridge
Stabilité des coefficients	✓	**Moins stable** que Ridge, (des coefficients sont nuls)
Utilisation optimale	Idéal lorsque les **variables sont corrélées** et toutes utiles	Idéal lorsqu'il y a **beaucoup de variables**, dont certaines non pertinentes
Quand utiliser ?	**Réduire la variance** sans éliminer de variables	**Simplifier le modèle** et à réduire la dimensionnalité

9. Extension aux GAM

Limites aux modèles linéaires

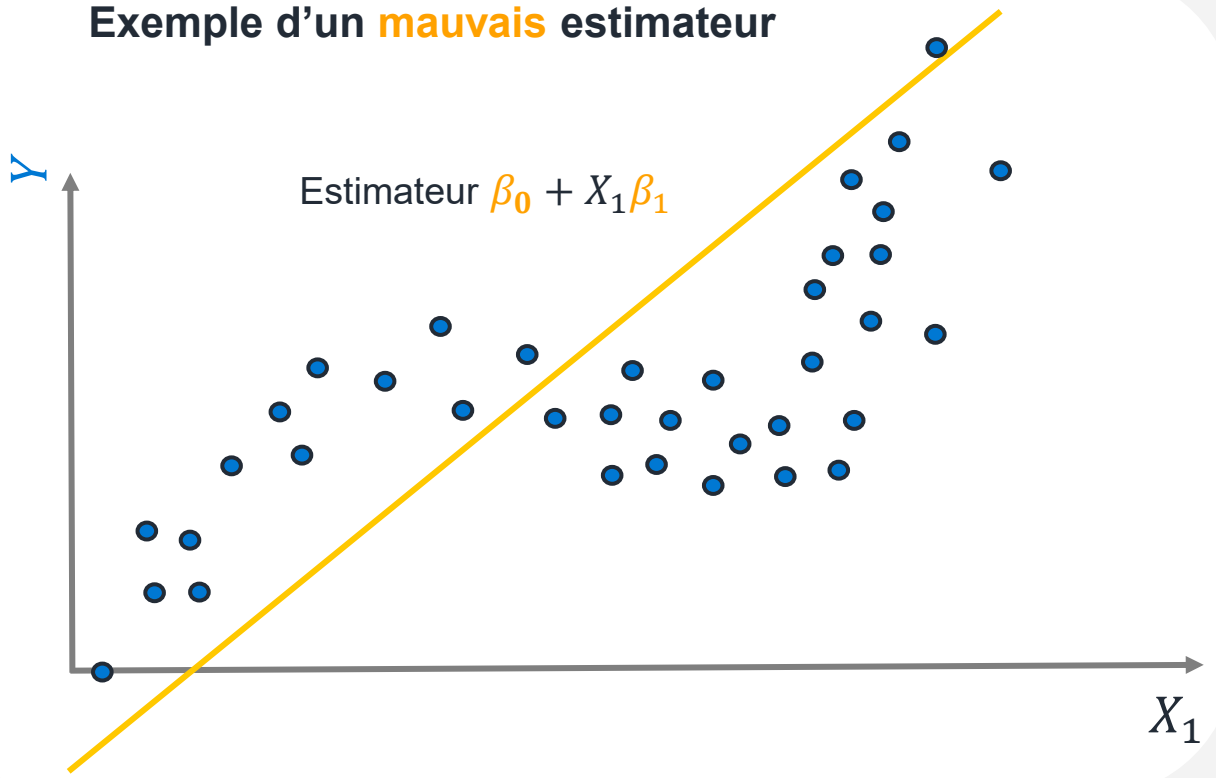
Cas relation non-linéaire entre variables explicatives et variable cible

Exemple illustratif

- **Modèle Linéaire Généralisé (GLM)**

$$g(\mathbb{E}[\,Y \mid \boldsymbol{X}\,]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

Exemple d'un mauvais estimateur



Limites aux modèles linéaires

Cas relation non-linéaire entre variables explicatives et variable cible

Exemple illustratif

Modèle Linéaire Généralisé (GLM)

$$g(\mathbb{E}[\,Y \mid \boldsymbol{X}\,]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

Modèle Additif Généralisé (GAM)

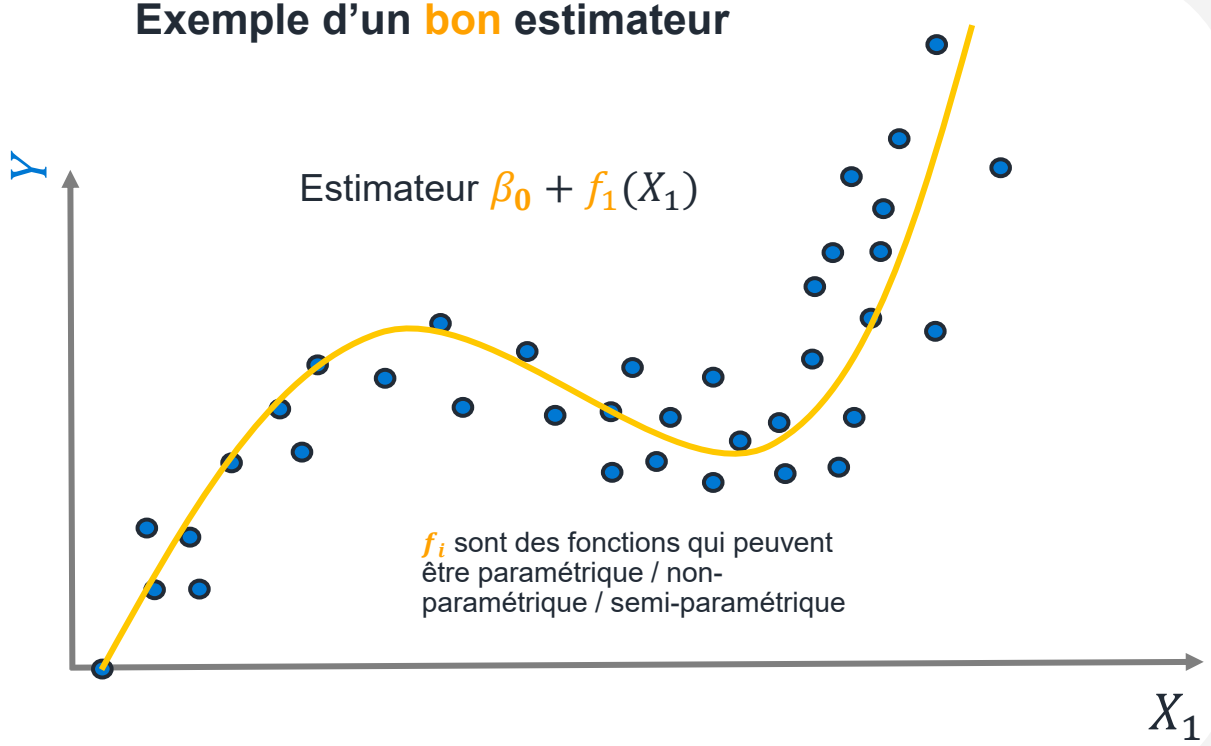
$$g(\mathbb{E}[\,Y \mid \boldsymbol{X}\,]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_d(X_d)$$

Les **GAM** (Modèles Additifs Généralisés) sont **une extension des GLM** qui permettent une plus grande flexibilité dans la modélisation des relations entre $\boldsymbol{X}$ et Y. Contrairement aux modèles GLM qui sont linéaires entre $g(\mathbb{E}[\,Y \mid \boldsymbol{X}\,])$ et $\boldsymbol{X}$, les GAM permettent à chaque variable explicative d'avoir une **relation non linéaire avec la réponse**.

Exemple de fonction :

$$f(x) = \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

Exemple d'un bon estimateur



Milliman

ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES

Merci

François HU

Francois.hu@milliman.com