Statistiques avancées — Régression
Cours 3 : Modèles linéaires généralisés

21 Septembre 2021

Introduction à la famille exponentielle

- La famille exponentielle est une **famille paramétrique** de distribution de probabilité, qui s'écrivent sous la même forme.
- Une variable aléatoire Y suit une distribution de la famille exponentiel si sa densité f s'écrit sous la forme :

$$f(Y) = h(Y) \exp\left(Y\theta^\star - g(\theta^\star)\right),$$

où $\theta^\star$ est un paramètre inconnu, et g et h sont des fonctions déterministes connues ; g est appelée la **fonction de lien** et h la **fonction de base**.

Premiers exemples

- **Distribution Gaussienne** : $Y \sim \mathcal{N}(\mu, 1)$ fait partie de la famille exponentielle, avec $\theta^\star = \mu$, $g = \frac{\mu^2}{2}$ et $h(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$.

- **Distribution de Bernoulli** : $Y \sim \mathcal{B}(p)$ fait partie de la famille exponentielle, avec $\theta^\star = \log\left(\frac{p}{1-p}\right)$, $g = \log(1 + \mathrm{e}^{\theta^\star})$ et $h(x) = 1$.

- **Distribution de Poisson** : $Y \sim \mathcal{P}(\lambda)$ fait partie de la famille exponentielle, avec $\theta^\star = \log \lambda$, $g = \exp(\theta^\star)$ et $h(x) = \frac{1}{y!}$.

Fonction génératrice des moments

Une propriété fondamentale de la famille exponentielle est que la fonction g donne les moments de la distributions :

$$\mathbb{E}_{\theta^\star}[Y] = g'(\theta^\star),$$
$$\mathrm{Var}_{\theta^\star}[Y] = g''(\theta^\star).$$

Exercice

Vérifier les deux relations ci-dessus pour les exemples des lois Gaussiennes, de Bernoulli et de Poisson.

Estimation du paramètre $\theta^\star$

- On s'intéresse à l'estimation du paramètre inconnu $\theta^\star$ par maximum de vraisemblance, sous l'hypothèse que $\theta_1^\star = \ldots = \theta_n^\star$, i.e. les Y_i sont i.i.d.

- La première étape est d'écrire la fonction de vraisemblance négative associée à l'échantillon, prend comme argument un paramètre $\theta \in \mathbb{R}$. Notons que la fonction de base $h(Y)$ qui ne dépend pas de θ disparait de la formule.

$$\begin{aligned} \log \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log(f(Y_i)) = \sum_{i=1}^{n} (Y_{ij}\theta - g(\theta)) \\ &= \theta \sum_{i=1}^{n} Y_{ij} - n g(\theta). \end{aligned} \tag{1}$$

Estimateur du maximum de vraisemblance $\hat{\theta}$

- Pour calculer l'estimateur du maximum de vraisemblance $\hat{\theta}$, on regarde les conditions d'optimalité du premier ordre qui donnent $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\theta} = 0$. On obtient :

$$\frac{\mathrm{d}\mathcal{L}(\theta)}{\mathrm{d}\theta} = \sum_{i=1}^{n} Y_{ij} - ng'(\theta),$$

Lorsque la fonction g' est inversible, on obtient

$$\hat{\theta} = (g')^{-1}\left(n^{-1}\sum_{i=1}^{n} Y_{ij}\right).$$

Exercice

Vérifier que g' est inversible pour les trois exemples cités plus hauts (Gaussien, Bernoulli, Poisson), et calculer les estimateurs associés.

Modèles linéaires généralisés

- Comme leur nom l'indique, les Modèles Linéaires Généralisés (GLM) généralisent le principe de la régression linéaire au-delà du modèle Gaussien.

- Les GLM sont utiles en particulier lorsque la variable réponse Y n'est pas continue (i.e. binaire, discrète, etc.) ou dévie fortement d'un modèle Gaussien.

Modèle général

- Dans le GLM, on ne suppose plus les Y_i i.i.d. mais on suppose que

$$Y_i \sim \mathrm{Exp}_{h,g}(\theta_i^\star),$$

  c'est-à-dire que chaque observation i a son propre paramètre $\theta_i^\star$.

- **Note** : Sans hypothèse supplémentaire, le modèle est surparamétré, i.e. il y a autant de paramètres que de données, et on ne peut espérer les estimer.

- On fait l'hypothèse supplémentaire d'un modèle linéaire. Soient $(X_i)_{i=1}^n$ des vecteurs de prédicteurs, avec $X_i \in \mathbb{R}^p$. On suppose la relation linéaire suivante entre X_i et $\theta_i^\star$:

$$\theta_i^\star = \beta_0 + \sum_{i=1}^{p} \beta_j X_{ij},$$

  où $\beta \in \mathbb{R}^{p+1}$ est le vecteur des coefficients de régression inconnus.

Modèle général

- Dans le modèle GLM, la moyenne de la$i$-ème réponse Y_i est donnée par

$$g'(\theta_i^\star) = g'\left(\beta_0 + \sum_{i=1}^{p} \beta_j X_{ij}\right).$$

- Étant donné un estimateur $\hat{\beta}$, on va en général prédire Y_i par sa moyenne, i.e.

$$\hat{Y}_i = g'\left(\hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_j X_{ij}\right) = \tilde{X}_i \hat{\beta},$$

avec $\tilde{X}_i = (1, X_i^\top)$.

Fonction de log-vraisemblance

- La fonction de log-vraisemblance associée à l'échantillon $(X_i, Y_i)_{1 \leq i \leq n}$ s'applique à un paramètre $\beta \in \mathbb{R}^{p+1}$ et s'écrit :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[Y_i \left(\tilde{X}_i \hat{\beta} \right) - g(\tilde{X}_i \hat{\beta}) \right]. \tag{2}$$

- Pour calculer l'estimateur du MLE $\hat{\beta}$, en admettant la concavité de $\mathcal{L}(\beta)$, on résoud l'équation

$$\nabla_\beta \mathcal{L}(\hat{\beta}) = 0.$$

Exercice

Montrer la concavite de $\beta \mapsto \mathcal{L}(\beta)$.

Calcul du MLE

- Pour $0 \leq j \leq p$, on a

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \left[Y_i \tilde{X}_{ij} - \tilde{X}_{ij} g'(\tilde{X}_i \hat{\beta}) \right].$$

- Le système $\frac{\partial \mathcal{L}}{\partial \beta_j} = 0$ pour tout j n'a en général pas de solution close. On utilise des méthodes de descente comme la descente de gradient ou la méthode de Newton-Raphson.

Exercice

Écrire une itération de l'algorithme de descente de gradient.

Newton—Raphson (Iteratively Reweighted Least Squares, IRLS)

Pour calculer (approcher) l'estimateur au maximum de vraisemblance, on utilise un algorithme de type Newton-Raphson. A l'étape k, on note $\hat{\beta}^k$ la solution courante. On approxime $-\frac{1}{n}\log\mathcal{L} = \ell_n$ par une fonction quadratique :

$$\ell_n(\hat{\beta}^k + h) = \ell_n(\hat{\beta}^k) + \nabla\ell_n(\hat{\beta}^k)^\top h + \frac{1}{2}h^\top\nabla^2\ell_n(\hat{\beta}^k)h$$

puis on minimise cette approximation pour obtenir h^* et on pose

$$\hat{\beta}^{k+1} = \hat{\beta}^k + h^*$$

puis on itère.

Exercice

Comprendre sur la régression de Poisson le nom "IRLS".

Loi asymptotique des estimateurs

On note $I(\beta) = -\mathbb{E}[\nabla^2 \ell_n]$

Consistence et normalité asymptotique

Sous certaines conditions (cf. Fahrmeir and Kaufman - 1985), on peut montrer que, pour tout vrai paramètre β,

1. $|\hat{\beta} - \beta| \to 0$
2. $\hat{\beta}$ est asymptotiquement gaussien, i.e. on a la convergence en loi suivante :

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}(0, I(\beta)^{-1}),$$

3. et

$$\sqrt{n} I(\hat{\beta})^{1/2}(\hat{\beta} - \beta) \to \mathcal{N}(0, Id).$$

C'est en particulier vrai pour les modèles considérés (à fonction de lien canonique) quand les covariables sont bornées.

Jeu de données Coronary Heart Disease (South Africa)

Échantillon d'hommes dans une région à haut risque de maladies cardiaques (Western Cape, Afrique du Sud).

- `sbp` pression artérielle systolique
- `tobacco` tabac cumulé (kg)
- `ldl` lipoprotéine de basse densité, mauvais cholestérol
- `famhist` antécédents familiaux de maladies cardiaques
- `typea` comportement de type A
- `obesity` obésité
- `alcohol` consommation actuelle d'alcool
- `age` âge
- `chd` réponse

On va modéliser l'apparition de CHD comme une série de tirages à pile ou face avec une probabilité de succès qui dépend des covariables ci-dessus.

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45

$n = 462$ patients, 160 cas (cdh $= 1$) et 302 controles

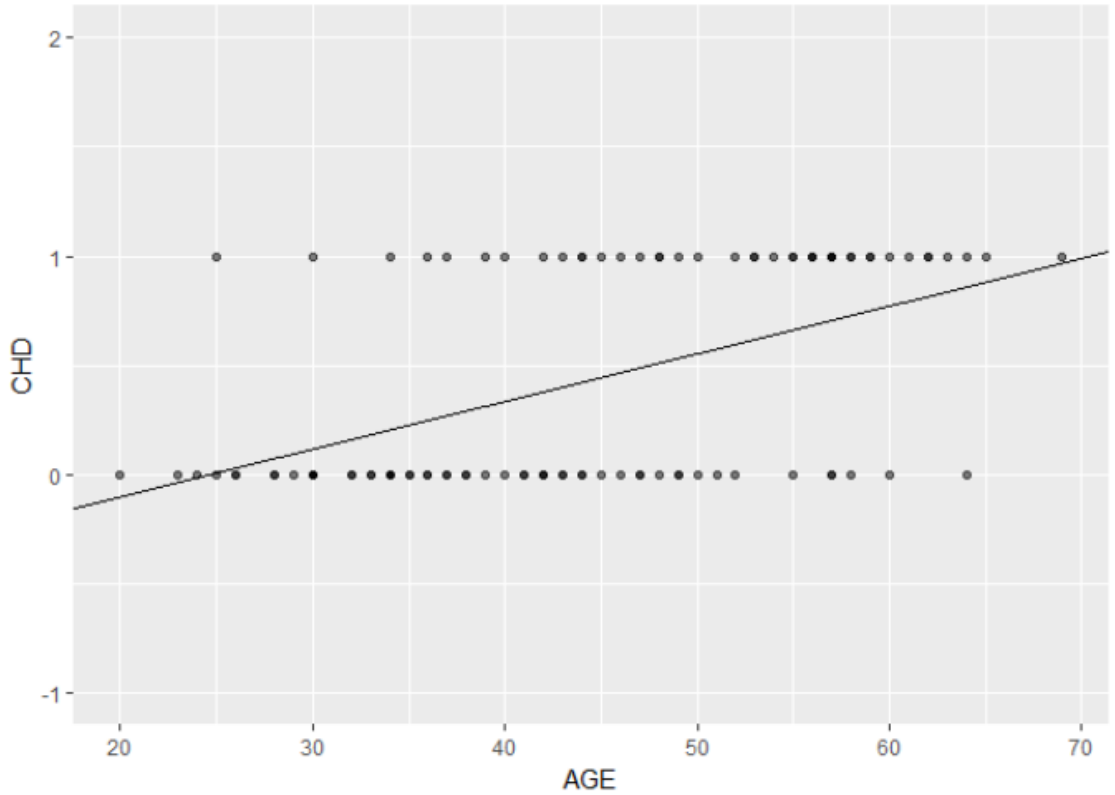
Questions liées aux données

- **Analyse** : Comprendre quels facteurs dans cet ensemble de données sont liés à la maladie (chd)
  - Importance de l'effet
  - Effet positif ou négatif
  - Significativité

- **Prédiction** : Prédire, pour un nouveau patient, le risque de déclarer la maladie.
  - Qualité de la prédiction
  - Interprétabilité avec un modèle parcimonieux

Échec de la régression linéaire

Peut-on utiliser la régression linéaire ?



Distribution de Bernoulli

- Données : réponse Y et prédicteurs $X \in \mathbb{R}^p$.
- Objectif : Prédire la probabilité que Y soit 1 ou 0 sachant la valeur de X.
- Loi de Bernoulli $\mathcal{B}(p)$ sur $\{0, 1\}$ telle que

$$Y \sim \mathcal{B}(p) \Leftrightarrow \begin{cases} \mathbb{P}(Y = 1) = p \\ \mathbb{P}(Y = 0) = 1 - p \end{cases}$$

Loi de Bernoulli conditionnelle

$$\begin{aligned} Y_i | X = x_i &\sim \mathcal{B}(p_i) \\ Y_i | X = x_i &\sim \mathcal{B}(\mathbb{P}(Y_i = 1 | X = x_i)) \end{aligned} \tag{3}$$

- On modélise la probabilité d'avoir une maladie connaissant les caractéristiques X
- C'est un jeu de pile ou face. Cependant, la probabilité de réussite sera différente d'une personne à l'autre en fonction de leurs covariables.

Modèle logistique

Modèle probabiliste

- $Y_i|X = i \sim \mathcal{B}(p_i)$
- $\mathbb{E}(Y_i|X = i) = p_i = \frac{\mathrm{e}^{\eta_\mathrm{i}}}{1+\mathrm{e}^{\eta_\mathrm{i}}}$
- On suppose la linéarite : $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$

Régression logistique, fonction en S

$$\mathbb{E}(Y_i|X = i) = p_i = \frac{\mathrm{e}^{\eta_\mathrm{i}}}{1 + \mathrm{e}^{\eta_\mathrm{i}}}$$

$$p_\beta(x_i) = \frac{\mathrm{e}^{\mathrm{x_i^t}\beta}}{1 + \mathrm{e}^{\mathrm{x_i^t}\beta}}$$

Outils pour l'interprétation

- On appelle "odds-ratio" le rapport des cotes $\frac{p_i}{1-p_i}$. Ici, c'est le rapport pour l'individu i de la probabilité d'avoir une CHD sur la probabilité de ne pas en avoir.
- On appelle le "log odds ratio" la quantité $\log\left(\frac{p_i}{1-p_i}\right) = \eta_i$.

$$\eta_i = \left(\frac{\mathbb{P}(Y_i = 1|X = x_i)}{\mathbb{P}(Y_i = 0|X = x_i)}\right) = \sum_{j=1}^{p} \beta_j x_{ij}$$

- Lorsque x_{ij} augmente de 1, la probabilité $\mathbb{P}(Y_i = 1|X = x_i)$ est multipliée par $\mathrm{e}^{\beta_\mathrm{j}}$.
- La fonction de lien g est la fonction logit $g : t \mapsto \log\left(\frac{t}{1-t}\right)$.

Estimateur du maximum de vraisemblance

Fonction de vraisemblance

La fonction de vraisemblance du modèle est définie par :

$$L_n(y_1, \dots, y_n, \beta) = \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i)$$

que l'on notera $L_n(\beta)$ par souci de simplicité.

On peut écrire cette formule comme une fonction du paramètre β :

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i) = \prod_{i=1}^{n} p_\beta(x_i)^{y_i} (1 - p_\beta(x_i))^{1-y_i}$$

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i) = \prod_{i=1}^{n} g^{-1}(x_i^t \beta)^{y_i} (1 - g^{-1}(x_i^t \beta))^{1-y_i},$$

où g est la fonction de lien logit.

Estimateur du maximum de vraisemblance

On continue le calcul :

$$\begin{aligned} L_n(\beta) &= \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\mathrm{x_i^t}\beta}}{1+\mathrm{e}^{\mathrm{x_i^t}\beta}} \right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{\mathrm{x_i^t}\beta}} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\mathrm{x_i^t}\beta \mathrm{y_i}}}{1+\mathrm{e}^{\mathrm{x_i^t}\beta}} \right) \end{aligned} \tag{4}$$

Log-vraisemblance

$$\log(L_n(\beta)) = \sum_{i=1}^{n} \left(y_i x_i^t \beta - \log(1+\mathrm{e}^{x_i^t \beta}) \right).$$

Calcul de l'estimateur : fonction de score

Pour calculer le minimiseur de la vraisemblance négative, on calcule les conditions d'optimalité du premier ordre (fonction de score).

$$S(\beta) = \nabla \log(L_n(\beta)) = \left(\frac{\partial \log(L_n(\beta))}{\partial \beta_0}(\beta), \dots, \frac{\partial \log(L_n(\beta))}{\partial \beta_p}(\beta) \right)$$

$$\begin{aligned} \frac{\partial \log(L_n(\beta))}{\partial \beta_j}(\beta) &= \sum_{i=1}^{n} \left(y_i x_{ij} - \frac{x_{ij} e^{x_i^t \beta}}{1 + e^{x_i^t \beta}} \right) \\ &= \sum_{i=1}^{n} x_{ij} \left(y_i - p_\beta(x_i) \right) \end{aligned} \tag{5}$$

Les conditions d'optimalité donnent :

$$S(\beta) = X^\top (Y - P_\beta) = 0.$$

Calcul du MLE

Malheureusement...

- En régression linéaire nous avons une forme close pour le MLE $\hat{\beta}$:

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

- En régression logistique ce n'est plus le cas, il n'y pas de forme close.

Algorithmes d'optimisation

Il va falloir utiliser des méthodes numériques pour calculer le MLE

- Descente de gradient, méthode de Newton, descente par coordonnées, etc.
- ... À voir dans le cours d'optimisation de la semaine 4 (P. Ablin).

Comportement asymptotique du MLE

Comme en régression linéaire, on dispose de nombreux résultats théoriques pour l'inférence.

Théorème

1. $\hat{\beta} \to_{p.s} \beta$ lorsque $n \to \infty$
2. $\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}(0, I(\beta)^{-1})$ où $I(\beta)$ est la matrice d'information de Fisher

$$I(\beta)_{k,l} = -\mathbb{E}\left(\frac{\partial^2 \log L_n}{\partial\beta_k \partial\beta_l L_n(\beta)}\right)$$

Théorème

$$(\hat{\beta} - \beta)^t n I(\beta)(\hat{\beta} - \beta) \to \chi^2_p$$

Intervalles de confiance asymptotiques

Distributions asymptotiques

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \to \chi_1^2,$$

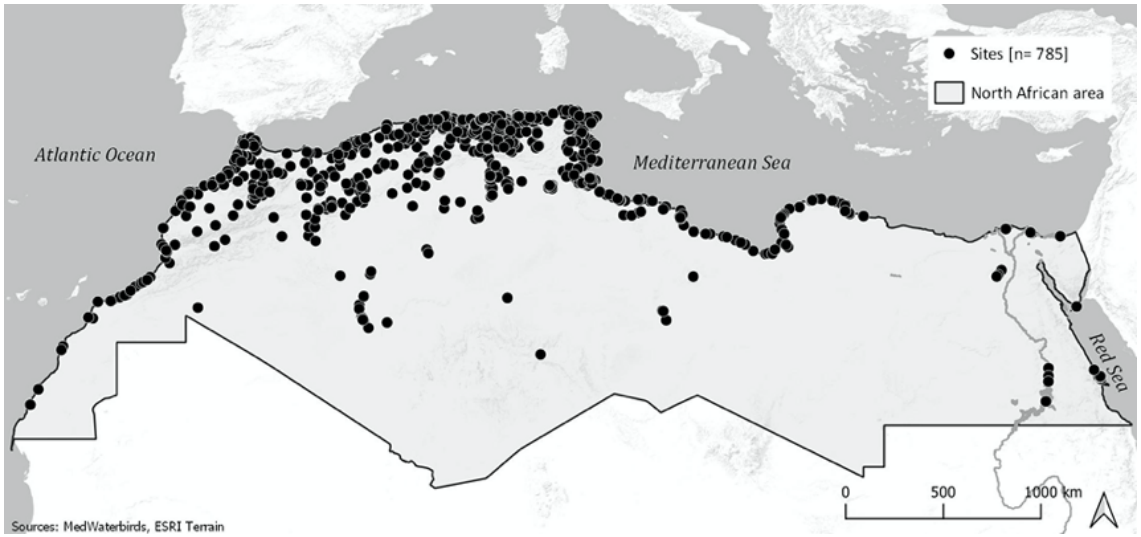
$$\frac{(\hat{\beta}_j - \beta_j)}{\hat{\sigma}_j} \to \mathcal{N}(0,1).$$

Intervalles de confiance asymptotique

$$I_{1-\alpha}(\beta_j) = \left[\hat{\beta} - u_{1-\alpha/2}\hat{\sigma}_j; \hat{\beta} + u_{1-\alpha/2}\hat{\sigma}_j\right].$$

Données de comptage : surveillance d'espèces sauvages

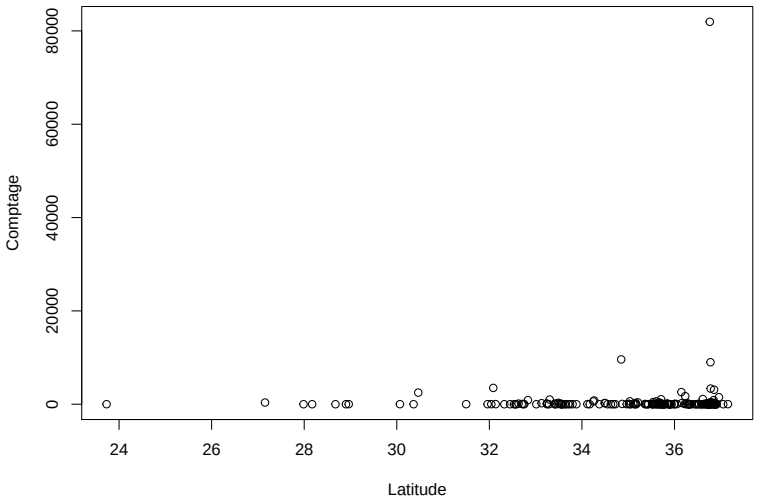
- Surveillance d'oiseaux d'eau en Afrique du Nord : comptage d'espèces dans des sites écologiques.

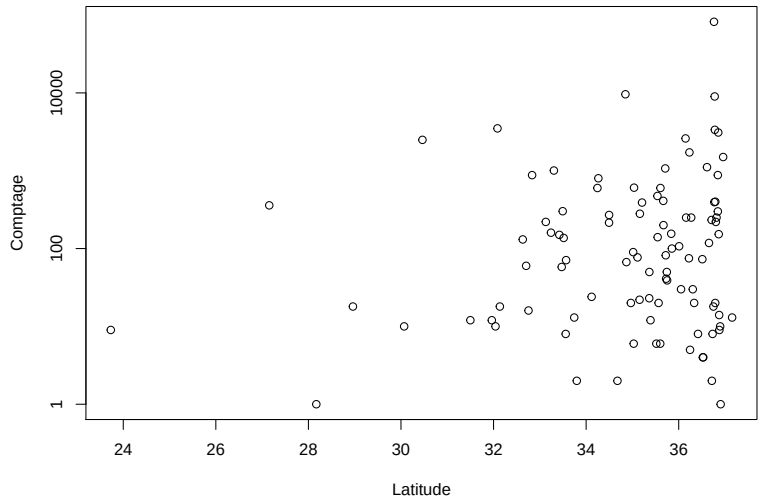


Données de comptage : surveillance d'espèces sauvages

- **Réponse** $Y_i \in \mathbb{N}$: nombre d'oiseaux compté au site i.
- **Prédicteurs** $X \in \mathbb{R}^p$: informations concernant les sites écologiques
  - `Latitude` : Latitude du site
  - `Longitude` : Longitude du site
  - `Altitude` : Altitude moyenne du site
  - `Distance town` : Distance à la ville la plus proche (en m.)
  - `Distance coast` : Distance à la côte (en m.)
  - `Surface eau` : Surface en eau du site (km2)
- **Objectif** : Prédire et expliquer le nombre d'oiseaux observés en fonction des prédicteurs géographiques.

Limites de la régression linéaire





- Données discrètes (entiers)
- Échelle non-linéaire (échelle normale à gauche, échelle log pour Y à droite)

Modèle de régression Poissonienne

- Modèle Poissonien sur la réponse (comptages)

$$\forall i \in \{1, \dots, n\}, \quad Y_i \sim \mathcal{P}(\lambda_i),$$

d'intensité $\mathbb{E}[Y_i] = \lambda_i$.

- Modèle 'log-linéaire' sur l'intensité λ_i

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

- La moyenne de Y dépend des covariables de la manière suivante :

$$\mathbb{E}[Y_i] = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right).$$

Fonction de vraisemblance du modèle Poissonien

- La fonction de log-vraisemblance associée à l'échantillon s'écrit

$$\mathcal{L}(\lambda_1, \dots, \lambda_n) = \sum_{i=1}^{n} \left(Y_i \log(\lambda_i) - \lambda_i \right).$$

- En utilisant le modèle log-linéaire $\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$ on peut réécrire $\mathcal{L}$ comme une fonction de $\beta = (\beta_0, \dots, \beta_p)$:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left(Y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \right) - \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \right) \right).$$

- On cherche le maximum de vraisemblance avec les conditions d'optimalité du premier ordre

$$\nabla_\beta \mathcal{L}(\hat{\beta}) = 0.$$

Maximum de vraisemblance

- Pour $1 \leq j \leq p$,

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left(Y_i X_{ij} - X_{ij} \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) \right) = 0.$$

- Comme pour la régression logistique, il n'y a pas de forme close. On doit calculer le maximum de vraisemblance numériquement (descente de gradient, etc.)

Motivations

Comme dans la régression linéaire multivariée, le MLE peut être un "mauvais" estimateurs lorsque :

- Le nombre de paramètres $p >> n$: MLE mal défini, infinité de solutions, grande variance asymptotique.
- Les données ne suivent pas exactement le modèle : présence d'outliers, corruptions dans les données.

Dans ce cas, on peut avoir recours à la pénalisation de la fonction de log-vraisemblance, par exemple avec la pénalité ridge (vue dans le cours d'hier).

Pénalité ridge en GLM

- La pénalité ridge s'ajoute à la fonction de log-vraisemblance négative :

$$\ell_{\texttt{ridge}}(\beta) = -\sum_{i=1}^{n} \left[Y_i \left(\tilde{X}_i \beta \right) - g(\tilde{X}_i \beta) \right] + \lambda \|\beta\|_2^2$$

- L'estimateur ridge en GLM est l'unique minimiseur de la log-vraisemblance négative pénalisée :

$$\hat{\beta}_{\texttt{ridge}} = \operatorname{argmin}_{\beta \in \mathbb{R}^{p+1}} \quad \ell_{\texttt{ridge}}(\beta).$$

- $\hat{\beta}_{\texttt{ridge}}$ est en général plus robuste que le MLE, peut donner de meilleures prédictions, est plus stables numériquement.

Motivations pour la pénalité Lasso

Dans certaines applications, le nombre de prédicteurs est si grand qu'on souhaite modifier la pénalité ridge pour forcer certains coefficients $\hat{\beta}_j$ à être nuls.

- Afin de rendre le modèle mieux spécifié (meilleures garanties statistiques)
- Afin d'améliorer l'interprétabilité du modèle.

Pour cela, on utilise la pénalité Lasso (vue en détail dans le cours de demain). La norme Euclidienne est remplacée par la norme 1 :

$$\ell_{\texttt{lasso}}(\beta) = -\sum_{i=1}^{n}\left[Y_i\left(\tilde{X}_i\beta\right) - g(\tilde{X}_i\beta)\right] + \lambda \underbrace{\|\beta\|_1}_{\sum_{j=0}^{p}|\beta_j|} .$$

Algorithmes pour les GLM pénalisés

Lorsqu'on introduit des pénalités il faut en général modifier les algorithmes d'optimisation pour calculer les estimateurs.

- En régression ridge, on peut conserver le principe de l'IRLS puisque la pénalité est quadratique.
- En régression Lasso, on ne peut pas appliquer cette méthode car la norme 1 $\|\beta\|_1$ n'est par dérivable et n'admet pas d'approximation quadratique. Alternatives :
  - Descente de gradient proximal (cf. cours de demain et cours d'optimisation).
  - Descente par coordonnées (cf. article de Friedman, Hastie Tibshirani).