⑥ Modèle linéaire généralisé

$Y_1, \dots, Y_n$ indépendants

On a vu que le modèle linéaire $Y_i = x_i^T \beta + \varepsilon_i$ (on suppose les x_i déterministes ici) permet de prédire simplement un label à valeurs réelles $Y_i \in \mathbb{R}$ à partir de vecteurs de features $X_i \in \mathbb{R}^d$ (ε_i est un bruit centré et $\mathrm{Var}(\varepsilon_i) = \sigma^2$)

Dans le modèle linéaire gaussien on avait supposé $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ de sorte que $Y_i \sim \mathcal{N}(x_i^T \beta, \sigma^2)$. Dans un modèle linéaire généralisé on va avoir un label Y_i distribué différent selon une loi appartenant à la famille de <u>modèles exponentiels</u>, l'espérance $\mu = \mathbb{E} Y$ dépend de $x \in \mathbb{R}^d$ à travers une <u>fonction de lien</u> g t.q.

$$g(\mathbb{E} Y_i) = x_i^T \beta \quad \text{ou bien} \quad \mu_i = \mathbb{E} Y_i = g^{-1}(x_i^T \beta)$$

On supposera g monotone et C^1 en général. Dans ce cadre, la variance s'écrit alors également $\mathrm{Var}\, Y_i = V(\mu_i) = V(g^{-1}(x_i^T \beta))$ pour une certaine fonction de variance V.

Def Un modèle linéaire généralisé pour des données $(x_1, Y_1), \dots, (x_n, Y_n)$ consiste en

1) Un modèle exponentiel pour $Y \in \mathbb{R}$
2) Un prédicteur linéaire $\eta_i = x_i^T \beta$ (i.e. $\beta \in \mathbb{R}^d$)
3) Une fonction de lien t.q. $g(\mathbb{E} Y_i) = x_i^T \beta$

1) On considère donc un modèle exponentiel pour Y : on considère une densité

$$f_\theta(y) = \exp\left(b(\theta) T(y) - A(\theta) \right) h(y)$$

où b, T, A et h sont des fonctions $\mathbb{R} \to \mathbb{R}$

C'est une petite généralisation du modèle exponentiel canonique vu avant : si on pose $\tilde{\theta} = b(\theta)$ (reparamétrisation)

on peut écrire $f_{\tilde{\theta}}(y) = \exp\left(\tilde{\theta}^T T(y) - \log Z(\tilde{\theta}) \right)$

avec $Z(\tilde{\theta}) = \exp\left(A(b^{-1}(\tilde{\theta})) \right)$ (si b inversible ...)

et mesure dominante $\tilde{\mu}(dy) = h(y)\, \mu(dy)$

On a vu que dans le modèle canonique où $b(\theta) = \theta$ et si $T(y) = y$ alors

$$\mathbb{E}_\theta Y = A'(\theta) \quad \text{et} \quad \mathrm{Var}_\theta Y = A''(\theta) \quad \text{où} \quad A(\theta) = \log Z(\theta).$$

2) L'idée est de lier ce modèle pour Y avec la valeur du prédicteur linéaire en posant simplement $\eta_i = x_i^T \beta$ (ce prédicteur linéaire $\in \mathbb{R}$ en général ...) et en supposant que $\mu_i = \mathbb{E} Y_i = g^{-1}(\eta_i) = g^{-1}(x_i^T \beta)$ pour une certaine fonction de lien g. Le lien entre le prédicteur linéaire et le modèle exponentiel se fait à travers l'espérance de Y et la fonction de lien g.

3) La fonction de lien g fait ce lien, un grand nombre de choix sont possibles. Cette fonction doit cependant aller de l'ensemble des valeurs que peut prendre $\mathbb{E} Y$ dans $\mathbb{R}$ ou entier. Il y a cependant un choix particulier de lien : la fonction de lien <u>canonique</u> définie lorsque l'on regarde le modèle exponentiel sous forme canonique.

Déf Si $Y \sim f_\theta$ avec $f_\theta(y) = \exp(\theta y - A(\theta))$ alors la fonction de lien canonique est la fonction g telle que $\theta = g(\mu)$ ie

$$\theta = g(\mathbb{E}_\theta Y)$$

Quand on choisit la fonction de lien canonique on a alors

$$\theta = g(\mu) = x^T\beta \quad \text{ie} \quad \theta_i = g(\mu_i) = x_i^T\beta$$

6.1 Modèle Bernoulli, régression logistique

Un exemple très important, modèle de régression logistique et certainement un des modèles les plus utilisés au monde.

Label binaire $Y_i \in \{0,1\}$ et features $x_i \in \mathbb{R}^d$. Par exemple clic ou non-clic sur une publicité sur internet... Comme $Y_i \in \{0,1\}$ forcément $Y_i \sim B(p_i)$ (Bernoulli de proba p_i)

Il faut faire un lien entre p_i et x_i : on commence par utiliser un prédicteur linéaire $\eta_i = x_i^T\beta$ (il faudra estimer $\beta \in \mathbb{R}^d$) et choisir une fonction de lien g telle que $g(\mathbb{E}Y_i) = g(p_i) = x_i^T\beta$

On a que $p_i \in [0,1]$ et $x_i^T\beta \in \mathbb{R}$ ($\beta \in \mathbb{R}^d$ peut prendre des valeurs arbitraires)

donc il faut choisir une fonction $g : [0,1) \to \mathbb{R}$ qui soit monotone et continue... De telles fonctions sont en fait des fonctions quantiles $g = F^{-1}$ où $F : \mathbb{R} \to [0,1)$ est la fonction de répartition d'une certaine loi. Par exemple si on choisit $F = \Phi$ = fonction de répartition de $N(0,1)$ on obtient la <u>régression probit</u>.

Quelle est la fonction de lien canonique pour ce modèle ? On a ($y \in \{0,1\}$)

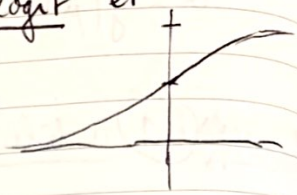
$$b_p(y) = p^y(1-p)^{1-y} = \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right)$$

et ici $\mathbb{E}Y = p$ et $\theta = \log\left(\frac{p}{1-p}\right)$ pour obtenir un modèle

canonique, donc la fonction de lien canonique est ici $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$

car alors $\theta = g(\mathbb{E}Y)$. Cette fonction s'appelle la fonction <u>logit</u> et

$\pi = \frac{1}{1+e^{-\theta}} = g^{-1}(\theta)$ s'appelle la fonction <u>sigmoïde</u>.



(une fonction de répartition bien particulière)

Ce modèle linéaire généralisé s'appelle la régression logistique. On suppose que

$$\mathbb{P}(Y=y \mid X=x) = \begin{cases} \sigma(\beta^T x) & y=1 \\ 1-\sigma(\beta^T x) & y=0 \end{cases}$$

dans ce modèle. On observe $(x_1, Y_1), \dots, (x_n, Y_n)$ et on veut estimer $\beta \in \mathbb{R}^d$.

Pour ce faire on utilise à nouveau l'estimateur du maximum de vraisemblance.

On va utiliser plutôt $Y \in \{-1, 1\}$ à la place de $Y \in \{0, 1\}$ (sans perte de généralité) car cela nous amènera à des formules plus élégantes.

On a

$$f_{Y|X=x}(y) = \begin{cases} \sigma(\beta^T x) & \text{si } y=1 \\ 1-\sigma(\beta^T x) & \text{si } y=-1 \end{cases}$$

mais σ est la fonction de répartition d'une loi symétrique donc $\sigma(-s) = 1 - \sigma(s)$

de sorte que $f_{Y|X=x}(y) = \sigma(y\beta^T x)$ et donc

la vraisemblance de $(Y_1, \dots, Y_n)$ avec Y_i indépendantes est donnée par

$$\beta \mapsto \prod_{i=1}^n \sigma(y_i x_i^T \beta)$$ et donc la moins log-vraisemblance

s'écrit

$$-\ell(\beta) = -\sum_{i=1}^n \log \sigma(y_i x_i^T \beta)$$

$$= \sum_{i=1}^n \log\left(1 + e^{-y_i^* x_i^T \beta}\right)$$

On note que $-\frac{1}{n}\ell(\beta) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+e^{-y_i x_i^T \beta}\right)$. Pour estimer β on cherche donc à minimiser la fonction $-\frac{1}{n}\ell(\beta)$. Cette fonction est "simple" à minimiser. Etudions là un peu: on pose $f(z) = \log(1+e^{-z})$ qui s'appelle la fonction de perte logistique. Elle s'interprète bien :

si $y_i x_i^T \beta \gg 0$ le prédicteur linéaire $x_i^T \beta$ est très positif et (resp négatif)
$y_i = 1$ (resp -1) $\Rightarrow$ β permet de bien prédire y_i et à l'inverse si

$y_i x_i^T \beta \ll 0$ alors le signe de $x_i^T \beta$ est fortement en désaccord avec celui de y_i $\rightarrow$ β se trompe fortement sur l'échantillon i.

En minimisant $\beta \mapsto -\frac{1}{n}\ell(\beta)$, on cherche un β qui, en moyenne sur l'échantillon a des petites "pertes" mesurées par la fonction logistique.

C'est un problème d'optimisation convexe: en effet $f'(z) = -\sigma(-z)$ $\left(= -\frac{1}{1+e^z}\right)$

et $f''(z) = \frac{e^z}{(1+e^z)^2} = \sigma(z)\sigma(-z) = \sigma(z)(1-\sigma(z)) \in [0, \frac{1}{4}]$

Donc $\nabla_\beta\left(-\frac{1}{n}\ell(\beta)\right) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\beta f(y_i x_i^T \beta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \sigma(-y_i x_i^T \beta)\, x_i$

et $\nabla_\beta^2\left(-\frac{1}{n}\ell(\beta)\right) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\beta^2 f(y_i x_i^T \beta) = \frac{1}{n}\sum_{i=1}^{n} \sigma(y_i x_i^T \beta)\left(1-\sigma(y_i x_i^T \beta)\right) x_i x_i^T$

et $0 \leq \sigma(1-\sigma) \leq \frac{1}{4}$ et $x_i x_i^T \succeq 0$ donc $0 \preceq \nabla_\beta^2\left(-\frac{1}{n}\ell(\beta)\right) \left(\preceq \frac{1}{4n}\sum x_i x_i^T\right)$

Hessienne $\succeq 0$ $\Rightarrow$ problème convexe.

Rem On peut rendre le problème strictement convexe (λ-fortement convexe) en considérant comme pour la régression linéaire moindres carrés une pénalisation ridge :

$$\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+e^{-y_i x_i^T \beta}\right) + \frac{\lambda}{2}\|\beta\|_2^2$$

(solution unique dans ce cas)

D'autres exemples de modèles classiques sont $\mathcal{P}(\lambda)$ = poisson et $\mathcal{N}(\mu, \sigma^2)$

Loi	Support	Lien canonique	Moyenne μ
$\mathcal{N}(\mu, \sigma^2)$	$\mathbb{R}$	$g(\mu) = \mu = x^T\beta$	$\mu = x^T\beta$
$\mathcal{P}(\lambda)$	$\mathbb{N}$	$g(\lambda) = \log\lambda = x^T\beta$	$\mu = e^{x^T\beta}$
$\mathcal{B}(p)$	$\{0,1\}$	$g(p) = \log\left(\frac{p}{1-p}\right)$	$p = \frac{1}{1+e^{-x^T\beta}}$
$\mathcal{E}(\lambda)$	$\mathbb{R}^+$	$g(\lambda) = -\frac{1}{\lambda}$	$\mu = -\frac{1}{x^T\beta}$

(6.2) Maximum de vraisemblance dans un GLM et TCL

On veut établir un TCL dans le modèle linéaire généralisé avec lien canonique, ie $Y \sim f_\beta$ où $f_\beta(y|x) = \exp\left(y x^T\beta - A(x^T\beta)\right) h(y)$

Rappel : pour la régression logistique $A(z) = -\log(1+e^z)$

$A'(z) = -\sigma(-z)$ $\quad A''(z) = -\sigma(z)(1-\sigma(z))$

La log vraisemblance s'écrit

$$\ell_n(\beta) = \sum_{i=1}^n \left\{ y_i x_i^T\beta - A(x_i^T\beta) - \log h(y_i) \right\}$$

et le score vaut

$$S_n(\beta) = \nabla \ell_n(\beta) = \sum_{i=1}^n \left\{ y_i - A'(x_i^T\beta) \right\} x_i$$

et l'information de Fisher

$$I_n(\beta) = \nabla^2 \ell_n(\beta) = -\sum_{i=1}^n A''(x_i^T\beta)\, x_i x_i^T .$$

On rappelle que $\mathbb{E}_\beta[S_n(\beta)] = 0$ et $\mathrm{cov}_\beta(S_n(\beta)) = I_n(\beta)$

On va avoir besoin de qq hypothèses qu'on va discuter rapidement :

On commence par poser $B_n(\delta) = \left\{ \beta \in \mathbb{R}^d : \left\| I_n(\beta_0)^{1/2} (\beta - \beta_0) \right\| \le \delta \right\}$

Hypothèses On suppose $Y_1, \dots, Y_n$ iid de densité f_{β_0} pour $\beta_0 \in \mathbb{R}^d$ = le "vrai" paramètre.

Hyp (I) $\lambda_{\min}\left(I_n(\beta_0)\right) \longrightarrow +\infty \qquad n \to +\infty$

Hyp (II) On pose $\tilde{I}_n(\beta) = I_n(\beta_0)^{-1/2} I_n(\beta) I_n(\beta_0)^{-1/2}$

matrice d'information de Fisher "normalisée" on suppose que

$$\forall \delta > 0 \qquad \max_{\beta \in B_n(\delta)} \left\| \tilde{I}_n(\beta) - Id \right\| \longrightarrow 0$$

On a que si $\beta \in B_n(\delta)$ alors $\|\beta - \beta_0\| \leq \dfrac{\delta}{\lambda_{\min}(I_n(\beta_0))} \to 0$

donc $B_n(\delta)$ se réduit à $\{\beta_0\}$ qd $n \to +\infty$ $(\forall \delta > 0)$

Donc sous (I)(II) est une hypothèse assez faible. De plus (II) équivaut à $\forall \varepsilon > 0$ $\forall u \in \mathbb{R}^d$ et $\forall \beta \in B_n(\delta)$ on a

$$\left| u^T (\tilde{I}_n(\beta) - Id) u \right| \leq \varepsilon \|u\|^2 \qquad \text{pour } n \text{ assez grand}$$

ie en posant $v = I_n^{-1/2}(\beta_0) u$ on a

$\forall \varepsilon > 0$ $\forall v \in \mathbb{R}^d$ $\forall \beta \in B_n(\delta)$

$$\left| v^T I_n(\beta) v - v^T I_n(\beta_0) \right| \leq \varepsilon \, v^T I_n(\beta_0) v \qquad n \text{ assez grand}$$

ce qui entraîne $(\varepsilon = 1/2)$ $I_n(\beta) \geq \frac{1}{2} I(\beta_0)$ $\forall \beta \in B_n(\delta)$ pour n assez grand

Thm Sous (I)+(II) on a pour $\hat{\beta}_n$ MLE (existe avec proba $\to 1$)

que $I_n(\beta_0)^{1/2} (\hat{\beta}_n - \beta_0) \rightsquigarrow N(0, Id)$

Dém Pour l'instant prenons $\hat{\beta}_n$ tq $S_n(\hat{\beta}_n) = 0$ (MLE si il existe)

et faisons un DL $\left(\text{"}\nabla \ell_n(\hat{\beta}_n)\right)$

$$S_n(\beta_0) = \underbrace{S_n(\hat\beta_n)}_{0} + \int_0^1 I_n(\beta_0 + t(\hat\beta_n - \beta_0))\,dt \quad (\hat\beta_n - \beta_0)$$

$$= I_n(\beta_0)^{1/2} \int_0^1 \tilde I_n(\beta_0 + t(\hat\beta_n - \beta_0))\,dt\; I_n(\beta_0)^{1/2} \quad (\hat\beta_n - \beta_0)$$

ie $$I_n(\beta_0)^{-1/2} S_n(\beta_0) = \int_0^1 \tilde I_n(\text{——})\,dt\; I_n(\beta_0)^{1/2} \quad (\hat\beta_n - \beta_0)$$

Donc si (A) $\int_0^1 \tilde I_n(\beta_0 + t(\hat\beta_n - \beta_0))\,dt \xrightarrow{\mathbb{P}} I_d$ et si

(B) $I_n(\beta_0)^{-1/2} S_n(\beta_0) \rightsquigarrow N(0, I_d)$ (C) $\hat\beta_n$ existe avec proba $\to 1$

alors on a $$I_n(\beta_0)^{1/2} (\hat\beta_n - \beta_0) \rightsquigarrow N(0, I_d).$$

Donc il faut montrer (A) + (B) + (C).

Preuve de (B)

Soit $\delta > 0$ et $u \in \mathbb{R}^d$, $\|u\| = 1$. On pose $\beta_n = \beta_0 + \delta I_n(\beta_0)^{-1/2} u$ de sorte que $\|\beta_n - \beta_0\|_{I_n(\beta_0)} = \delta$ ie $\beta_n \in \partial B_n(\delta)$ $(\subset B_n(\delta))$.

On fait un DL : on a pour un certain $\tilde\beta_n \in [\beta_n, \beta_0] = \{t\beta_n + (1-t)\beta_0 : t \in [0,1]\}$

$$l_n(\beta_n) = l_n(\beta_0) + \langle \beta_n - \beta_0, \nabla l_n(\beta_0)\rangle + \frac{1}{2} (\beta_n - \beta_0)^T \nabla l_n^2(\tilde\beta_n) (\beta_n - \beta_0)$$

$$= l_n(\beta_0) + \delta \langle u, I_n(\beta_0)^{-1/2} S_n(\beta_0)\rangle - \frac{\delta^2}{2} u^T \tilde I_n(\tilde\beta_n) u$$

et enfin $$L_n(\beta_n) = e^{\frac{\delta^2}{2} u^T \tilde I_n(\tilde\beta_n) u}\; e^{\delta u^T I_n(\beta_0)^{-1/2} S_n(\beta_0)}\; L_n(\beta_0)$$

où $L_n(\beta) = e^{l_n(\beta)}$ = vraisemblance du modèle en β. Donc en intégrant on obtient

$$\mathbb{E}_{\beta_n}\left[\exp\left(\frac{\delta^2}{2} u^T I_n(\tilde\beta_n) u\right)\right] = \mathbb{E}_{\beta_0}\left[\exp\left(\delta u^T I_n(\beta_0)^{-1/2} S_n(\beta_0)\right)\right]$$

On a $\beta_n \in B_n(\delta)$ donc (II) entraîne $\|\tilde{I}_n(\tilde{\beta}_n) - I_d\| \to 0$

donc $\tilde{\beta}_n \in B_n(\delta)$

et par continuité de l'exponentielle on obtient $\forall \varepsilon > 0 \quad \left| e^{\frac{\delta^2}{2} u^T \tilde{I}_n(\tilde{\beta}_n) u} - e^{\delta^2/2} \right| \leq \varepsilon$

pour n assez grand et donc $\mathbb{E}_{\beta_n} \exp\left(\frac{\delta^2}{2} u^T I_n(\tilde{\beta}_n) u\right) \to e^{\delta^2/2}$

et donc $\mathbb{E}_{\beta_0}\left[\exp\left(\delta u^T I_n(\beta_0)^{-1/2} S_n(\beta_0)\right)\right] \to e^{\delta^2/2}$ $n \to +\infty$

ce qui entraîne que $I_n(\delta_0)^{-1/2} S_n(\beta_0) \rightsquigarrow N(0, I_d)$

(la transformée de Laplace converge vers la transformée de Laplace de $N(0, I_d)$)

Preuve de (A) On veut montrer que

$$\left\| \int_0^1 \tilde{I}_n(\beta_0 + t(\hat{\beta}_n - \beta_0))\, dt - I_d \right\| \xrightarrow{P} 0$$

On remarque que $\beta_0 + t(\hat{\beta}_n - \beta_0) \in B_n(\delta)$ si $\hat{\beta}_n \in B_n(\delta)$

Il faut que l'on étudie $\hat{\beta}_n$ comme un maximum quand il existe. On rappelle que $I_n(\beta) = -\nabla^2 l_n(\beta) = -\sum_{i=1}^n A''(x_i^T \beta)\, x_i x_i^T < 0 \quad \forall \beta \in \mathbb{R}^d$

donc $l_n(\beta)$ est strictement concave sur $\mathbb{R}^d \Rightarrow$ il y a au plus un $\beta \in \mathbb{R}^d$ tq $S_n(\beta) = 0$ et c'est dans ce cas $\hat{\beta}_n$ = le MLE.

Si $l_n(\beta) - l_n(\beta_0) < 0 \quad \forall \beta \in \partial B_n(\delta)$ (pour un certain $\delta > 0$)

alors il y a un maximum dans $B_n(\delta)$ (car l_n bornée au bord de $B_n(\delta)$)

et dans ce cas, c'est $\hat{\beta}_n$. On a montré que

$$\forall \eta > 0 \ \exists \delta > 0 \text{ tq } P\left(l_n(\beta) - l_n(\beta_0) < 0 \ \forall \beta \in \partial B_n(\delta)\right) \geq 1 - \eta$$

Soit $\beta \in \partial B_n(\delta)$ de sorte que $u = \frac{1}{\delta} I_n(\beta_0)^{1/2}(\beta - \beta_0)$ vérifie $\|u\| = 1$

et on refait encore le DL

$$l_n(\beta) - l_n(\beta_0) = \delta \left\langle u, I_n(\beta_0)^{-1/2} S_n(\beta_0) \right\rangle - \frac{\delta^2}{2} u^T \tilde{I}_n(\tilde{\beta}_n) u \quad \text{pour } \tilde{\beta}_n \in [\beta_0, \beta_n]$$

Rem : $\max_{\|v\|=1} \langle v, I_n(\beta_0)^{-1/2} S_n(\beta_0) \rangle = \| I_n(\beta_0)^{-1/2} S_n(\beta_0) \|$ et

$\inf_{\|v\|=1} v^T I_n(\tilde{\beta}_n) v = \lambda_{min}(\tilde{I}_n(\tilde{\beta}_n))$ donc il suffit de montrer

$$\mathbb{P}\left[\| I_n(\beta_0)^{-1/2} S_n(\beta_0) \|^2 < \frac{\delta^2}{4} \lambda_{min}(\tilde{I}_n(\tilde{\beta}_n))^2 \right] \geq 1-\eta$$

mais (II) $\Rightarrow$ $I_n(\beta) \gtrsim c\, I_n(\beta_0)$ $\forall \beta \in B_n(\delta_0)$

donc $I_n(\beta_0)^{-1/2} I_n(\beta) I_n(\beta_0)^{-1/2} \gtrsim c\, I_d$

$\| \quad \tilde{I}_n(\beta)$

donc il suffit de vérifier que $\mathbb{P}\left[\| I_n(\beta_0)^{-1/2} S_n(\beta_0) \|^2 < \frac{(\delta c)^2}{4} \right] \geq 1-\eta$

Mais il suffit d'appliquer Markov : $\mathbb{P}\left[\| - \|^2 < \frac{(\delta c)^2}{4} \right] \geq 1 - \frac{4 \mathbb{E} \| I_n(\beta_0)^{-1/2} S_n(\beta_0) \|^2}{(\delta c)^2}$

et enfin $\mathbb{E} \| I_n(\beta_0)^{-1/2} S_n(\beta_0) \|^2$

$= \mathbb{E}\left[S_n(\beta_0)^T I_n(\beta_0)^{-1} S_n(\beta_0) \right] = \mathbb{E}\, tr(S^T I^{-1} S)$

$= \mathbb{E}\, tr(I^{-1} S S^T) = tr(I^{-1} \mathbb{E}(S S^T))$

$= tr\left(I_n^{-1}(\beta_0) \underbrace{\mathbb{E}\left[S_n(\beta_0) S_n(\beta_0)^T \right]}_{\| \quad \mathrm{Var}_{\beta_0}(S_n(\beta_0)) = I_n(\beta_0)} \right)$

$= d$ donc si on choisit $\delta = \sqrt{\frac{4d}{c^2 \eta}}$ et pour

n assez grand (tq $I_n(\beta) \gtrsim I_n(\beta_0)$ $\forall \beta \in B_n(\delta)$) on obtient bien

$$\mathbb{P}\left[\ell_n(\beta) - \ell_n(\beta_0) < 0 \quad \forall \beta \in \partial B_n(\delta) \right] \geq 1-\eta$$

et on remarque que sur cet événement $\hat{\beta}_n$ existe et $\hat{\beta}_n \in B_n(\delta)$

donc $\mathbb{P}\left[\hat{\beta}_n \in B_n(\delta)\right] \geq 1-\eta$

$$\mathbb{P}\left[\left\| I_n(\beta_0)^{1/2}(\hat{\beta}_n - \beta_0)\right\| \leq \delta\right] \geq 1-\eta .$$

On a aussi $\beta_0 + t(\hat{\beta}_n - \beta_0) \in B_n(\delta)$ si $\hat{\beta}_n \in B_n(\delta)$ donc en utilisant II on obtient $\left\| \tilde{I}_n(\beta_0 + t(\hat{\beta}_n - \beta_0)) - I_d \right\| \to 0$

et donc $\left\| \int_0^1 \tilde{I}_n(\beta_0 + t(\hat{\beta}_n - \beta_0))\, dt - I_d \right\| \to 0$ sur cet événement

on a donc montré que $\forall \varepsilon > 0$ et $\forall \eta > 0$ on peut trouver δ_η pour n assez grand

$$\mathbb{P}\left[\left\| \int_0^1 \tilde{I}_n(\text{---})\, dt - I_d \right\| > \varepsilon\right] \leq \eta$$

ce qui montre (A) □

Concernant les hypothèses : si $c_1 \leq A''(z) \leq c_2$ pour $z \in (H)$ $0 < c_1 < c_2$

et $(x_i)_{i \geq 1}$ et β_{η} tq $x_i^T \beta \in (H)$ ($\beta \in B$ compact)

alors $I_n(\beta_0) = \sum_{i=1}^n A''(x_i^T \beta_0)\, x_i x_i^T$

$$\geq \min_{i=1,\dots,n} A''(x_i^T \beta_0) \sum_{i=1}^n x_i x_i^T$$

si $\lambda_{\min}\left(\sum_{i=2}^n x_i x_i^T\right) \to +\infty$ alors $\lambda_{\min}\left(I_n(\beta_0)\right) \to +\infty$ et (II) est vérifiée si $1-\varepsilon \leq \dfrac{A''(x_i^T \beta)}{A''(x_i^T \beta_0)} \leq 1+\varepsilon$ $\forall i$

mais A C^∞ et (convexe) sur (H)

Ex log-reg $\dfrac{\sigma(x_i^T\beta)(1-\sigma(x_i^T\beta))}{\sigma(x_i^T\beta_0)(1-\sigma(x_i^T\beta_0))}$