Jackknife, bootstrap et cross-validation

Agathe Guilloux

Section 1

Introduction

But de l'inférence statistique

On a

- $\mathcal{X} = (X_1, \ldots, X_n)$ un échantillon i.i.d. de fonction de répartition F
- $\theta(F)$ une quantité d'intérêt, qui dépend de F
- $T(\mathcal{X})$ une statistique, estimateur de $\theta(F)$,

on voudrait connaître

- le biais : $\mathbb{E}_F(T(\mathcal{X})) - \theta(F)$
- la variance : $\mathbb{E}_F(T^2(\mathcal{X})) - \mathbb{E}_F^2(T(\mathcal{X}))$
- le MSE (EQM) : $\mathbb{E}_F\left((T(\mathcal{X}) - \theta(F))^2\right)$
- la loi de $T(\mathcal{X})$: $G(x) = \mathbb{P}_F(T(\mathcal{X}) \leq x)$, $\forall x$.
- etc

pour comparer des estimateurs, connaître leurs qualités, construire des intervalles de confiance...
Problème : toutes ses quantités dépendent de la loi F inconnue !

Ce que l'on sait

On a à disposition la fonction de répartition empirique des X_i.

Fonction de répartition empirique

Pour $\mathcal{X} = (X_1, \ldots, X_n)$, la fonction de répartition empirique est définie par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_(X_i \leq x), \ \forall x.$$

On va considérer les estimateurs par plug-in.

Principe de plug-in

Pour tout paramètre $\theta(F)$ et tout échantillon $\mathcal{X} = (X_1, \ldots, X_n)$, on considère l'estimateur par plug-in

$$T(\mathcal{X}) = \theta(F_n) = \hat{\theta}$$

$\rightarrow$ exemples : espérance, variance, médiane

Méthode de rééchantillonnage

Idée générale :

- Si on avait plusieurs échantillons $\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots,$
- on aurait plusieurs copies de $T(\mathcal{X})$: $T(\mathcal{X}^{(1)}), T(\mathcal{X}^{(2)}), \ldots$
- et on pourrait estimer sa loi et toutes les quantités.

Différentes méthodes :

- le jackknife - Quenouille (1949)[Que56], puis Tuckey (1958)[Tuk58]
- le bootstrap - Efron (1979)[Efr92]

Bibiographie

- Jackknife, bootstrap and other resamplings plans (1982). B. Efron [Efr82]
- An introduction to the bootstrap (1993). B. Efron et R. Tibshirani[ET94]
- Bootstrap methods and their applications (1997). A. Davidson et D. Hinkley [DH97]

Section 2

Partie 1 : Jackknife

Estimateur de Quenouille du biais

A partir de l'observation $\mathbf{x} = (x_1, \ldots, x_n)$, on construit les échantillons

$$\mathbf{x}^{(-i)} = (x_1, \ldots, x_{i-1}, x_i, \ldots)$$

de taille $n-1$. On note $F_n^{(-i)}$ la f.d.r empirique associée à chaque échantillon puis on calcule

$$\theta(F_n^{(-i)}) = \hat{\theta}^{(-i)}$$

suivant le principe du plug-in.

Estimateur de Quenouille du biais

Quenouille propose d'estimer le biais de $\theta(F_n) = \hat{\theta}$ par

$$\widehat{Biais} = (n-1)(\hat{\theta}^{(\cdot)} - \hat{\theta})$$

où

$$\hat{\theta}^{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}^{(-i)}.$$

Cas de la variance

Estimateur du jackknife corrigé du biais

On définit alors l'estimateur de $\theta(F)$ du jackknife corrigé du biais par

$$\tilde{\theta} = \hat{\theta} - \widehat{Biais} = \hat{\theta} - (n-1)(\hat{\theta}^{(\cdot)} - \hat{\theta}) = n\hat{\theta} - (n-1)\hat{\theta}^{(\cdot)}.$$

On peut montrer que si le biais de $\hat{\theta}$ est de la forme $\sum_{k\geq 1} a_k(F)/n^k = O(1/n)$ alors le biais de $\tilde{\theta}$ est de l'ordre de $1/n^2$.

$\rightarrow$ cas de la variance

Estimateur jackknife de la variance

Tuckey propose de considérer des “pseudo-values”

$$\tilde{\theta}_i = \hat{\theta} + (n-1)\big(\hat{\theta} - \hat{\theta}^{(-i)}\big).$$

On a la relation

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_i.$$

Estimateur du jackknife de la variance de $\hat{\theta}$

On définit

$$\widehat{\mathbb{V}(\hat{\theta})} = \text{"}\widehat{\mathbb{V}(\tilde{\theta})}\text{"} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\big(\tilde{\theta}_i - \tilde{\theta}\big)^2$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}\big(\hat{\theta}^{(-i)} - \hat{\theta}^{(\cdot)}\big)^2$$

Exercice

1. Exemple “law school data”. Appliquer la méthode de Quenouille pour obtenir une estimation du biais de la corrélation entre les scores “LSAT” et “GPA”.
2. Donner un estimateur jackknife de la variance de la corrélation pour le jeu de données “law school”.
3. Simuler $n = 10$ réalisations de $Z \sim \mathcal{E}(1)$. Calculer la médiane obtenue sur les réalisations puis estimer sa variance par jackknife. Recommencer en faisant grandir n et comparer à la variance limite.

Section 3

Partie 2 : Boostrap

Bootstrap d'Efron

A partir de l'observation $\mathbf{x} = (x_1, \ldots, x_n)$, on construit des échantillons

$$\mathcal{X}_1^* = (X_{1,1}^* = x_{m_1}, \ldots, X_{1,n}^* = x_{m_n})$$
$$\ldots$$
$$\mathcal{X}_b^* = (X_{b,1}^* = x_{m_{(b-1)n+1}}, \ldots, X_{b,n}^* = x_{m_{bn}})$$
$$\ldots$$

où les m_k ont été tirés aléatoirement et avec remise dans $\{1, \ldots, n\}$.

Loi des $X_{b,j}^*$ conditionnelle à $\mathcal{X}$

Conditionnellement $\mathcal{X}$, $X_{b,j}^*$ est une v.a. de fonction de répartition F_n, fonction de répartition empirique des $X_1, \ldots, X_n$.

Estimateurs du bootstrap classique

Soit un paramètre inconnu $\theta(F)$

- Monde réel
  - avec l'échantillon initial $\mathcal{X}$, on définit l'estimateur $\hat{\theta} = \theta(F_n) = T(\mathcal{X})$
  - on note G_n la f.d.r. inconnue de $\hat{\theta}$, qui dépend de F, inconnue
- Monde bootstrap
  - pour chaque échantillon bootstrapé $\mathcal{X}_b^*$, on définit l'estimateur $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$
  - conditionnellement à F_n, de loi G_n^* qui dépend de F_n
  - on estime G_n^* par

$$\hat{G}_{n,B}^*(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_(\hat{\theta}_b^* \leq t)$$

Exemples d'estimateurs bootstrap (1)

Estimation de la loi de $\hat{\theta}$

La f.d.r. G_n de $\hat{\theta}$ est définie pour $t \in \mathbb{R}$ par

$$G_n(t) = \int \mathbb{1}_(x \leq t) dG_n(x)$$

elle est estimée par (1ere approximation du bootstrap)

$$G_n^*(t) = \int \mathbb{1}_(x \leq t) dG_n^*(x) = \mathbb{P}_{F_n}(\hat{\theta}_b^* \leq t)$$

puis (2ieme approximation du bootstrap)

$$G_n^*(t) = \int \mathbb{1}_(x \leq t) d\hat{G}_{B,n}^*(x).$$

Exemples d'estimateurs bootstrap (2)

Estimation du biais de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F(T(\mathcal{X})) - \theta(F) = \int x dG_n(x) - \theta(F)$$

est estimé (1ere approximation) par

$$\int x dG_n^*(x) - \theta(F_n)$$

puis (2ieme approximation) par

$$\int x d\hat{G}_{B,n}^*(x) - \theta(F_n) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* - \theta(F_n).$$

Exemples d'estimateurs bootstrap (3)

Estimation de la variance de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F((T(\mathcal{X}) - \mathbb{E}_F(T(\mathcal{X}))^2) = \int (x - \int x dG_n)^2 dG_n(x)$$

est estimé (1ere approximation) par

$$\int (x - \int x dG_n^*)^2 dG_n^*(x)$$

puis (2ieme approximation) par

$$\int (x - \int x d\hat{G}_{B,n}^*)^2 d\hat{G}_{B,n}^*(x) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*)^2.$$

etc

Exercice

1. Donner un estimateur bootstrap de la variance de la corrélation pour le jeu de données “law school”.
2. Simuler $n = 10$ réalisations de $Z \sim \mathcal{E}(1)$. Calculer la médiane obtenue sur les réalisations puis estimer sa variance par bootstrap. Recommencer en faisant grandir n et comparer à la variance limite.

Comparer les résultats à ceux obtenus par jackknife.

Eléments de validation asymptotique du bootstrap (1)

Le bootstrap fait deux approximations

$$G_n \overset{(1)}{\to} G_n^* \overset{(2)}{\to} \hat{G}_{B,n}^*.$$

Pour contrôler la deuxième approximation, on utilise une borne de Dvoretsky-Kiefer-Wolfowitz.

Borne DKW, DKW (1956) - Massart (1990)

Si $Z_1, \ldots, Z_N$ est un échantillon i.i.d. de f.d.r. H et H_N est la f.d.r. empirique associée alors

$$\mathbb{P}\left(\sqrt{N}\sup_{x\in\mathbb{R}}|H_N(x) - H(x)| > \epsilon\right) \leq 2\exp(-2\epsilon^2).$$

Application pour le choix de B :
Si on veut que $\sup_{x\in\mathbb{R}}|\hat{G}_{B,n}^*(x) - G_n^*(x)| \leq 0\ .02$ avec une probabilité plus grande que $0\ .05$, comment choisir B ?

Eléments de validation asymptotique du bootstrap (2)

La première approximation est contrôlée par les développements d'Edgeworth. Si $\hat{\theta}$ est asymptotiquement normal :

$$S = \sqrt{n}\frac{\hat{\theta} - \theta}{\sigma(F)} \xrightarrow{\mathcal{F}} \mathcal{N}(0,1)$$

avec quelques conditions supplémentaires, on peut montrer que G_n admet un développement d'Edgeworth

$$\mathbb{P}(S \leq x) = G_n(\theta + \sigma x/\sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} p(x)\phi(x) + \mathcal{O}(\frac{1}{n}).$$

Dans le monde bootstrap, on peut montrer que si

$$S^* = \sqrt{n}\frac{\hat{\theta}^* - \hat{\theta}}{\sigma(F)}$$

on a

$$\mathbb{P}_{F_n}(S^* \leq x) = G_n^*(\theta + \sigma x/\sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} \hat{p}(x)\phi(x) + \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

Eléments de validation asymptotique du bootstrap (3)

Le point clé est que $p(x) - \hat{p}(x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n^{1/2}})$. Un simple calcul montre alors :

- Approximation gaussienne

$$\mathbb{P}(S \leq x) - \Phi(x) = \frac{1}{n^{1/2}} p(x)\phi(x) + \mathcal{O}(\frac{1}{n}) = \mathcal{O}(\frac{1}{n^{1/2}})$$

- Approximation bootstrap

$$\mathbb{P}(S \leq x) - \mathbb{P}_{F_n}(S^* \leq x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

Simulations

Exercice

Comparer sur simulation

- la fonction de répartition de la médiane empirique (approchée par Monte Carlo)
- la fonction de répartition de l'approximation asymptotique gaussienne
- la fonction de répartition empirique des pseudo-values du jackknife et
- la fonction de répartition empirique des estimateurs bootstrapés

dans le cas d'un n-échantillon i.i.d. de loi $\mathcal{E}(1)$ ($n = 10$, puis $n = 50$).

Intervalle de confiance du bootstrap basique

On définit les statistiques d'ordre des estimateurs bootstrap

$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}$$

IC du bootstrap basique

$$\widehat{IC}^*_{basic}(1-\alpha) = \left[2\hat{\theta} - \hat{\theta}^*_{(\lceil B(1-\alpha/2)\rceil)}, 2\hat{\theta} - \hat{\theta}^*_{(\lceil B\alpha/2\rceil)}\right]$$

Intervalle de confiance du percentile

S'il existe une fonction h telle que la loi de $h(T)$ est symétrique autour de $h(\theta)$.

IC du percentile

$$\widehat{IC}^*_{perc}(1-\alpha) = \left[\hat{\theta}^*_{(\lceil B\alpha/2 \rceil)}, \hat{\theta}^*_{(\lceil B(1-\alpha/2) \rceil)}\right]$$

Intervalle de confiance du t-boostrap

Si on connaît un estimateur $\hat{\sigma} = \sigma(F_n) = \sigma(\mathcal{X})$ de la variance asymptotique $\sigma^2(F)$ de $T = \hat{\theta}$, on considère la statistique studentisée

$$S = \sqrt{n}\frac{\hat{\theta} - \theta}{\sigma(F_n)}$$

et ses versions boostrapées

$$S_b^* = \sqrt{n}\frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)}.$$

IC du t-boostrap

$$\widehat{IC}_{tboot}^*(1-\alpha) = \left[\hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S^*_{(\lceil B(1-\alpha/2)\rceil)}, \hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S^*_{(\lceil B(\alpha/2)\rceil)}\right]$$

Test via l'intervalle de confiance du t-bootstrap

On considère le problème de test de $\mathcal{H}_0 : \theta = \theta_0$ v.s. $\mathcal{H}_1 : \theta \neq \theta_0$. On peut faire ce test par bootstrap en comparant la statistique de test

$$\bar{S} = |\sqrt{n}\frac{\hat{\theta} - \theta_0}{\sigma(F_n)}|$$

aux statistiques bootstrapées

$$\bar{S}_b^* = |\sqrt{n}\frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)}|.$$

On définit alors la p-value boostrapée

$$\hat{p}_B = \frac{\#\{b : \bar{S}_b^* > \bar{S}\} + 1}{B + 1}$$

Si l'estimateur de la variance n'est pas disponible, on peut utiliser la statistique $|\hat{\theta} - \theta_0|$ et $|\hat{\theta}_b^* - \hat{\theta}|$.

Exercice

Voir le sujet du TP sur ma page.

Section 4

Modèles de régression

Introduction

Problème de la régression

On considère l'échantillon i.i.d. $\mathcal{S} = \Big((Y_1 , X_1) , \ldots , (Y_n , X_n) \Big)$ avec

- $Y_i(\Omega) \subset \mathbb{R}$
- $X_i(\Omega) \subset \mathbb{R}^p$.

On veut estimer $\mathbb{E}(Y_i | X_i) = g(X_i)$. On se donne un estimateur $\hat{g} = \hat{g}_{\mathcal{S}}$

Remarque : la fonction g est entièrement déterminée par la loi conditionnelle de Y_i sachant X_i et donc par la loi de (Y_i , X_i).

Exemples

Régression linéaire

$$Y_i = X_i\beta + \epsilon_i$$

donc $g(x) = x\beta$ et si $\epsilon \sim \mathcal{L}_\epsilon(0\,, \sigma^2)$ alors $Y_i|X_i \sim \mathcal{L}_\epsilon(X_i\beta\,, \sigma^2)$. On se donne $\hat{g}(x) = x\hat{\beta}$ avec $\hat{\beta}$ estimateur des moindres carrés de β.

Régression logistique

$Y_i(\Omega) = \{0\,, 1\}$ et

$$\mathbb{E}(Y_i|X_i) = \mathbb{P}(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} = \pi(X_i\beta)$$

alors $Y_i|X_i \sim \mathcal{L}(\pi(X_i\beta)\,, \pi(X_i\beta)(1 - \pi(X_i\beta))$. On se donne $\hat{g}(x) = \pi(x\hat{\beta})$ avec $\hat{\beta}$ estimateur au maximum de vraisemblance de β.

Rééchantillonnage des cas ("cases sampling")

Cases sampling

Puisque les (Y_i, X_i) sont supposés i.i.d.,

1. on échantillonne aléatoirement avec remise dans l'échantillon $\mathcal{S} = \Big((Y_1, X_1), \ldots, (Y_n, X_n)\Big)$ pour obtenir

$$\mathcal{S}_1^* = \Big((Y_{1,1}^*, X_{1,1}^*), \ldots, (Y_{1,n}^*, X_{1,n}^*)\Big)$$

$$\ldots$$

$$\mathcal{S}_B^* = \Big((Y_{B,1}^*, X_{B,1}^*), \ldots, (Y_{B,n}^*, X_{B,n}^*)\Big)$$

2. on calcule sur chaque échantillon bootstrappé $\hat{g}_{\mathcal{S}_b^*}$.

Rééchantillonnage des erreurs ("errors sampling") dans le modèle linéaire

Dans le modèle de régression linéaire, on définit les résidus par

$$e_i = Y_i - X_i\hat{\beta} \ \ \forall i = 1, \ldots, n.$$

On montre que

$$\mathbb{E}(e_i) = 0 \text{ et } \mathbb{V}(e_i) = (1 - H_{ii})\sigma^2.$$

On définit alors les résidus studentisés

$$e_i^S = \frac{e_i}{\sqrt{1 - H_{ii}}\hat{\sigma}_{-i}}$$

avec $H = X(X^\top X)^{-1}X^\top$.
Les e_i^S sont censés être proches en loi des e_i/σ^2. Si, par exemple, les ϵ_i sont gaussiens, on peut montrer

$$e_i^* \sim \mathcal{T}(n - p - 1).$$

Errors sampling

1. On calcule les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ à partir de $\mathcal{S}$, puis les résidus studentisés $e_1^S, \ldots, e_n^S$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^S, \ldots, e_n^S)$ pour obtenir

$$(e_{1,1}^{S,*}, \ldots, e_{1,n}^{S,*}) \ldots (e_{B,1}^{S,*}, \ldots, e_{B,n}^{S,*})$$

3. on reconstruit pour chaque b et chaque i

$$Y_{b,i}^* = X_i \hat{\beta} + \hat{\sigma} e_{b,i}^{S,*}$$

4. on calcule dans chaque échantillon bootstrapé des estimateurs de β et σ^2.

Rééchantillonnage des erreurs ("errors sampling") dans le modèle logistique

Dans le modèle logistique, il n'y a pas d'erreur ϵ. On définit cependant

- les résidus de Pearson

$$e_i^P = \frac{Y_i - \hat{\pi}(X^i)}{\sqrt{\hat{\pi}(X^i)(1-\hat{\pi}(X^i))}}.$$

- les résidus de déviance

$$\begin{aligned} e_i^D &= \sqrt{-2\log(\hat{\pi}(X^i)} \text{ si } Y_i = 1 \\ &= -\sqrt{-2\log(1-\hat{\pi}(X^i))} \text{ si } Y_i = 0. \end{aligned}$$

Ils jouent le même rôle que les résidus studentisés e_i^S du modèle linéaire. En particulier, on peut écrire "à la louche"

$$Y_i = \pi(X_i\beta) + Y_i - \pi(X_i\beta) = \pi(X_i\beta) + ``\epsilon_i'' \ \ \forall i = 1, \ldots, n$$

avec ϵ_i à valeurs sur $\{-\pi(X_i\beta), 1-\pi(X_i\beta)\}$, qui vérifient

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}(\epsilon_i) = \pi(X_i\beta)(1-\pi(X_i\beta).$$

Errors sampling

1. On calcule les estimateurs $\hat{\beta}$ à partir de $\mathcal{S}$, puis les résidus de Pearson (ou de déviance) $e_1^P, \ldots, e_n^P$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^P, \ldots, e_n^P)$ pour obtenir

$$(e_{1,1}^{P,*}, \ldots, e_{1,n}^{P,*}) \ldots (e_{B,1}^{P,*}, \ldots, e_{B,n}^{P,*})$$

3. on calcule pour chaque b et chaque i

$$\zeta_{b,i} = \pi(X_i\hat{\beta}) + \sqrt{\pi(X_i\hat{\beta})(1-\pi(X_i\hat{\beta})}e_{b,i}^{P,*}.$$

4. si $\zeta_{b,i} < 1/2$, on fixe $Y_{b,i}^* = 0$ et si $\zeta_{b,i} \geq 1/2$, on fixe $Y_{b,i}^* = 1$
5. on calcule dans chaque échantillon bootstrapés des estimateurs de β.

NB : cet algorithme s'étend simplement à tous les modèles linéaires généralisés.

Implémentation

1. Appliquer l'algorithme "cases sampling" sur les jeux de données "mtcars" et "urine". Les descriptions sont disponibles sur ma page.
2. Appliquer l'algorithme "errors sampling" sur le jeux de données "mtcars".
3. Donner via les deux algorithmes des IC par bootstrap basique et percentile pour les coefficients de la régression et les comparer aux IC classiques (par approximation gaussienne)

Application aux tests par bootstrap : errors sampling

Grâce à l'algorithme "errors sampling", on peut construire des tests par bootstrap. On note, pour chaque i, $X_i = (X_i^a, X_i^b)$ avec $X_i^a \in \mathbb{R}^{p^a}$ et $X_i^b \in \mathbb{R}^{p^b}$ de sorte que

$$X_i\beta = X_i^a\gamma + X_i^b\delta.$$

Supposons qu'on veut tester $H_0 : \delta = \vec{0}$ v.s $\bar{H_0}$. Les statistiques de tests à considérer sont

- la statistique de Fisher dans le modèle de régression linéaire

$$F = \frac{(n-p)\left(\|Y - X^a\hat{\gamma}\|^2 - \|Y - X\hat{\beta}\|^2\right)}{\|Y - X\hat{\beta}\|^2}$$

- la statistique du rapport de vraisemblance dans le modèle logistique

$$\Lambda = -2\left[\log\left(\mathcal{V}(\hat{\gamma})\right) - \log\left(\mathcal{V}(\hat{\beta})\right)\right].$$

On a besoin pour garantir un niveau α au test (ou pour calculer une p-value) de connaître la loi de F et Λ sous H_0.

Boostrap sous H_0

1. On calcule l'estimateur $\hat{\gamma}$ à partir de $\mathcal{S}$ dans le modèle sous H_0, puis les résidus studentisés (ou de Pearson ou de déviance) $e_1^{S,a}, \ldots, e_n^{S,a}$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^{S,a}, \ldots, e_n^{S,a})$ pour obtenir

$$(e_{1,1}^{S,a,*}, \ldots, e_{1,n}^{S,a,*}) \ldots (e_{B,1}^{S,a,*}, \ldots, e_{B,n}^{S,a,*})$$

3. on calcule pour chaque b et chaque i

$$Y_{b,i}^* = ???$$

4. ???

Bootstrap sous H_0

1. On calcule les estimateurs $\hat{\gamma}$ et $\hat{\sigma}^{a,2}$ à partir de $\mathcal{S}$ dans le modèle sous H_0, puis les résidus studentisés (ou de Pearson ou de déviance) $e_1^{S,a}, \ldots, e_n^{S,a}$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^{S,a}, \ldots, e_n^{S,a})$ pour obtenir

$$(e_{1,1}^{S,a,*}, \ldots, e_{1,n}^{S,a,*}) \ldots (e_{B,1}^{S,a,*}, \ldots, e_{B,n}^{S,a,*})$$

3. on calcule pour chaque b et chaque i

$$Y_{b,i}^* = X_i^a \hat{\gamma} + \hat{\sigma}^{a,2} e_{b,i}^{S,a,*}$$

4. on calcule les valeurs bootstrapées $F_1^{*,H_0}, \ldots, F_B^{*,H_0}$

Exercice

- Dans le cas de l'error sampling, comment approximer la p-value ?
- Comment construire le test dans le cas du case sampling

Implémentation

1. Adapter l'algorithme à la régression logistique.
2. Appliquer les algorithmes "errors sampling" et "case sampling" pour tester la nullité des coefficients sur les jeux de données "mtcars" et "urine".

Exercice : prédiction dans le modèle linéaire

- On dispose d'un échantillon $\mathcal{S} = \Big((Y_1 , X_1) , \ldots , (Y_n , X_n) \Big)$ (échantillon d'apprentissage)
- On dispose des variables explicatives X_+ pour un nouvel individu. On note Y_+ la valeur non-observée de sa réponse.
- On note l'erreur de prédiction

$$\text{Ep}(+) = \hat{Y}_+ - Y_+ = X_+\hat{\beta} - Y_+ = X_+\hat{\beta} - (X_+\beta + \epsilon_+)$$

1. En notant G la f.d.r. (inconnue) de $\text{Ep}(+)$, donner un IP exact pour Y_+.
2. Proposer des versions bootstrapées de $\text{Ep}(+)$
3. Donner un IP par bootstrap basique pour Y_+.

Section 5

Erreur de généralisation, cross-validation et bootstrap

Vrai modèle, sélection de modèle

Modèles, vrai modèle

On se donne une famille de modèles $\mathcal{M}$, par exemple $\mathcal{M} = \mathcal{P}\{1, \ldots, p\}$. On suppose qu'il existe un vrai modèle $m^* \in \mathcal{M}$ tel que :

$$Y = X^{(m^*)}\beta^{(m^*)} + \epsilon^* \quad \text{avec } \epsilon_i^* \, i.i.d. (\sigma^{m^*})^2 I_n).$$

avec ϵ_i i.i.d., $\mathbb{E}(\epsilon_i) = 0$ et $\mathbb{V}(\epsilon_i) = \sigma^2$.
sélection de modèle : on veut retrouver m^*.

Estimation dans le modèle m

Dans le modèle m, on note $|m|$ le nombre de covariables qu'il contient et

$$\hat{\beta}^{(m)} = \left((X^{(m)})^\top X^{(m)}\right)^{-1} (X^{(m)})^\top Y$$

$$\hat{Y}^{(m)} = X^{(m)}\hat{\beta}^{(m)}$$

$$\widehat{(\sigma^m)^2} = \frac{\|Y - \hat{Y}^{(m)}\|_2^2}{n - |m|}$$

Moindres carré, risque quadratique et validation interne

Le risque quadratique de $\hat{Y}^{(m)}$ pour l'estimation de $X^*\beta^*$ est donné par

$$\mathbb{E}(\|\hat{Y}^{(m)} - X^*\beta^*\|^2) = \underbrace{\mathbb{E}(\|\hat{Y}^{(m)} - X^{(m)}\beta^{(m)}\|^2)}_{\text{variance}} + \underbrace{\|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2}_{\text{biais}^2}$$
$$= \sigma^2|m| + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2$$

où $X^{(m)}\beta^{(m)}$ est la projection de $X^*\beta^*$ sur vect$(X^{(m)})$. Pour l'espérance de l'erreur de prédiction (erreur apparente), on montre que

$$\mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) = (n - |m|)\sigma^2 + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2.$$

Finalement

$$\begin{aligned}\mathbb{E}(\|\hat{Y}^{(m)} - X^*\beta^*\|^2) &= \sigma^2|m| - (n - |m|)\sigma^2 + \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) \\ &= 2\sigma^2|m| + \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) - n\sigma^2.\end{aligned}$$

Cp de Mallows

On choisit $\hat{m}_{Cp} \in \mathcal{M}$ tel que :

$$\hat{m}_{Cp} = \underset{m \in \mathcal{M}}{\operatorname{argmin}}\, Cp(m),$$

avec

$$Cp(m) = \frac{\widehat{(\sigma^m)^2}}{\widehat{(\sigma^{m_{tot}})^2}} + 2\frac{|m|}{n}$$

Validation externe

Si on avait à disposition d'autres données, on aurait

- des données d'apprentissage (training, learning set)
  $\mathcal{S}_L = \{(Y_1, X_i), \ldots, (Y_n, X_n)\}$
- des données de validation, de test (testing, validation set)
  $\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \ldots, (Y_{+,n'}, X_{+,n'})\}$ avec $Y_+ = X_+\beta^* + \epsilon_+$ **ET** ϵ et ϵ_+ indépendants.

On choisirait alors le modèle qui minimise l'erreur de généralisation.

Generalization error, erreur de généralisation

$$\mathbb{E}(\| Y_+ - \hat{Y}_+^{(m)} \|^2) = \mathbb{E}(\| Y_+ - X_+\hat{\beta}^{(m)} \|^2) = n'\sigma^2 + |m|\sigma^2 + \| X^{(m)}\beta^{(m)} - X^*\beta^* \|^2.$$

Théoriquement, on choisit le même modèle qu'avec le risque quadratique :

$$\mathbb{E}(\| \hat{Y}^{(m)} - X^*\beta^* \|^2) = \sigma^2|m| + \| X^{(m)}\beta^{(m)} - X^*\beta^* \|^2$$

Excès d'erreur

Si on minimisait

$$\mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) = (n - |m|)\sigma^2 + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2.$$

on choisirait toujours le plus grand modèle.

Excess error, excès d'erreur

On définit l'excès d'erreur comme

$$\mathbb{E}(\|Y_+ - \hat{Y}^{(m)}\|^2) - \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2)$$

Estimation de l'erreur de généralisation et de l'excès d'erreur

On estime l'erreur de généralisation $\mathbb{E}(\| Y_+ - \hat{Y}_+^{(m)} \|^2) = \mathbb{E}(\| Y_+ - X_+ \hat{\beta}^{(m)} \|^2)$ à partir de l'échantillon $\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \ldots, (Y_{+,n'}, X_{+,n'})\}$ par

$$\frac{1}{n'} \sum_{i=1}^{n'} (Y_{+,i} - X_{+,i} \hat{\beta}^{(m)})^2,$$

où $\hat{\beta}^{(m)}$ a été calculé sur l'échantillon d'apprentissage $\mathcal{S}_L$ et dans le modèle m.

En pratique

Même en l'absence de données de validation (situation fr ?uente en pratique), on peut vouloir créer des données qui "ressemblent" à des données de test pour appliquer ce qui précède. Il y a deux grandes méthodes

- la cross-validation
- le bootstrap

Leave-one-out (jackknife)

Chaque observation joue à tour de rôle le rôle d'échantillon de validation.

Estimation de l'erreur de généralisation par leave-one-out

$$\mathbb{E}(\| \widehat{Y_+ - \hat{Y}_+^{(m)}} \|^2)_{loo} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \hat{\beta}_{(-i)}^{(m)})^2,$$

où $\hat{\beta}_{(-i)}^{(m)}$ a été calculé sur l'échantillon $\mathcal{S}_L \backslash (Y_i, X_i)$ et dans le modèle m.

K-fold cross-validation

On découpe l'échantillon initial en K sous-ensembles pour obtenir la partition $\mathcal{S}_L = \mathcal{S}_{L,1} \cup \ldots \cup \mathcal{S}_{L,K}$. Dans le cas, où $n = Kn_K$, on tire aléatoirement et sans remise dans $\mathcal{S}_L$ pour former les $\mathcal{S}_{L,k}$.

Estimation de l'erreur de généralisation par K-fold cross-validation

$$\widehat{eg(m)}_{Kfold-cv} = \mathbb{E}(\| Y_+ - \widehat{\hat{Y}_+^{(m)}} \|^2)_{Kfold-cv} = \frac{1}{n_K K} \sum_{k=1}^{K} \sum_{i=1}^{n_K} (Y_{k,i} - X_{k,i} \hat{\beta}_{(-k)}^{(m)})^2,$$

où $\hat{\beta}_{(-k)}^{(m)}$ a été calculé sur l'échantillon $\mathcal{S}_L \backslash \mathcal{S}_{L,k}$ et dans le modèle m. On peut préférer l'ajustement

$$\widehat{eg(m)}_{A-Kfold-cv} = \mathbb{E}(\| Y_+ - \widehat{\hat{Y}_+^{(m)}} \|^2)_{A-Kfold-cv} = \mathbb{E}(\| Y_+ - \widehat{\hat{Y}_+^{(m)}} \|^2)_{Kfold-cv}$$
$$+ \frac{1}{n} \| Y - X\hat{\beta} \|^2 - \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} (Y_i - X_i \hat{\beta}_{(-k)}^{(m)})^2.$$

Estimation de l'excès d'erreur

Une estimation bootstrap de l'excès d'erreur

$$ee(m) = \mathbb{E}\big(\|Y_+ - \hat{Y}^{(m)}\|^2 - \|\hat{Y}^{(m)} - Y\|^2\big)$$

est

$$\widehat{ee(m)}^*_{boot} = \frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}\Big((Y_i - X_i\hat{\beta}^*_b)^2 - (Y^*_{b,i} - X^*_{b,i}\hat{\beta}^*_b)^2\Big).$$

Estimation de l'erreur de généralisation par bootstrap

On obtient alors un estimateur par bootstrap de l'erreur de généralisation

$$\widehat{eg(m)}_{boot} = \widehat{ee(m)}^*_{boot} + \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{(m)} - Y_i)^2$$

Section 6

TP

Comparaisons des méthodes de sélection de modèles

Partie 1 (une simulation) :

1. Simuler dans un modèle linéaire un échantillon d'apprentissage avec $n = 50$, $p = 5$
   - X de taille $(n \times p)$ de coordonnées i.i.d. de loi uniforme sur $[-0.5, 0.5]$
   - $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \epsilon$
   - avec $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 2$, , $\beta_3 = 0$, $\beta_4 = 0$,, $\beta_5 = 0$
   - ϵ_i i.i.d. de loi $\mathcal{N}(0, 1)$.
2. Pour une simulation, classer les variables explicatives $X^1, \ldots, X^5$ par ordre décroissant de lien avec Y (via les pvalues des tests de Student ou les corrélations). Au lieu des $2^5 = 32$ modèles à parcourir, on parcourra les modèles à 1, puis 2, puis 3 variables, etc, en ajoutant les variables par ordre décroissant de lien avec Y. On a donc 6 modèles à comparer pour tous les critères.
3. Choisir un modèle via le Cp de Mallows
4. Simuler dans un modèle linéaire un échantillon de test/validation avec $n = 50$, $p = 5$. Choisir le modèle via l'estimation de l'erreur de généralisation.
5. Choisir un modèle via la cross-validation leave-one-out
6. Choisir un modèle via la cross-validation K-fold (pour K=5, K=2)
7. Choisir un modèle via le bootstrap ($B = 100$) avec des tailles d'échantillons bootstrappés $n' = 25$ et $n' = 50$.

Comparaisons des méthodes de sélection de modèles

Partie 2 (Monte Carlo) :

1. Reproduire l'expérience $M = 100$ fois (ou 500 si votre ordinateur le permet)
2. Calculer pour chacun des critères de sélection la proportion de fois (sur les M expériences) où il choisit un modèle avec le bon nombre de variables.

Partie 3 (Influence des valeurs de p et β) :

1. Que se passe-t-il si $\beta_1 = \beta_2 = 0\,.5$? Pourquoi ?
2. Que se passe-t-il si $p = 10$, $\beta_0 = \beta_1 = \beta_2 = 2$, et , $\beta_3 = \ldots = \beta_{10} = 0$? Pourquoi ?

References I


Anthony Christopher Davison and David Victor Hinkley, *Bootstrap methods and their application*, vol. 1, Cambridge university press, 1997.


Bradley Efron, *The jackknife, the bootstrap and other resampling plans*, SIAM, 1982.


_____, *Bootstrap methods : another look at the jackknife*, Breakthroughs in Statistics, Springer, 1992, pp. 569–593.


Bradley Efron and Robert J Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.


Maurice H Quenouille, *Notes on bias in estimation*, Biometrika **43** (1956), no. 3/4, 353–360.


John W Tukey, *Bias and confidence in not-quite large samples*, Annals of Mathematical Statistics, vol. 29, INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545, 1958, pp. 614–614.