TP2 : Bootstrap et subsampling

ENSAE 3ème année, 2017

Applications du bootstrap et autres techniques de rééchantillonnage (cours de M. Roquain)

Exercice 0 (Création de votre fichier)

Créer votre fichier `nom_prenom_TP2.R`. N'oubliez pas d'effectuer des commentaires dans votre fichier (que vous utilisiez un notebook ou pas).

Exercice 1 (Loi du min)

Soit $\mathcal{X}_n = (X_1, \dots, X_n)$ un échantillon de variables i.i.d. suivant la distribution $P = U(\theta, \theta+1)$, pour $\theta \in \mathbb{R}$ ($n = 1000$). On considère l'estimateur $\hat{\theta}_n = X_{(1)} = \min_{1 \le i \le n}\{X_i\}$ de θ. On s'intéresse à la distribution

$$L_n(P) = \mathcal{L}(n(\hat{\theta}_n - \theta)).$$

A. Bootstrap

1) A l'aide d'un qqplot et de simulations, vérifier le théorème limite

$$n(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{E}(1).$$

2) Soit $\mathcal{X}_n^\star = (X_1^\star, \dots, X_n^\star)$ i.i.d. de loi $\hat{P}_n$ (la distribution empirique de $\mathcal{X}_n$). Utiliser la library `boot` pour approcher la distribution

$$\mathcal{L}(n(\hat{\theta}_n^\star - \hat{\theta}_n) \mid \mathcal{X}_n).$$

Pour $B = 10000$, combien y-a t'il d'éléments différents dans votre approximation bootstrap ?

3) Comparer cette distribution à la loi $\mathcal{E}(1)$ à l'aide d'un qqplot. Conclure.

B. Subsampling

Pour $b \in \{1, \dots, n-1\}$ donné, on propose d'approcher $L_n(P) = \mathcal{L}(n(\hat{\theta}_n - \theta))$ par $L_{n,b}(\hat{P}_n)$, la loi uniforme sur

$$\left\{ b\left(\hat{\theta}_b(X_j, j \in S) - \hat{\theta}_n\right), S \subset \{1, \dots, n\}, |S| = b \right\}.$$

1) Construire une fonction `subsample` qui prend en entrée $\mathcal{X}_n$ et b et qui renvoie une valeur de $b\left(\hat{\theta}_b(X_j, j \in S) - \hat{\theta}_n\right)$, pour un S tiré uniformément parmi les sous ensembles de $\{1, \dots, n\}$ de taille b. On pourra utiliser la fonction `sample`.

2) Calculer B réalisations i.i.d. de $L_{n,b}(\hat{P}_n)$. On pourra utiliser la fonction `apply`.

3) Comparer la loi $L_{n,b}(\hat{P}_n)$ à la loi $\mathcal{E}(1)$ pour $b = 50$ à l'aide d'un qqplot.

4) Que vaut l'estimateur jackknife de la loi $L_n(P)$? Commenter.

Exercice 2 (Regression linéaire)

Télécharger le fichier *law82.dat* sur ma page http://etienne.roquain.free.fr/teaching.html, puis attacher les variables `GPA` et `LSAT` de la façon suivante :

```
data=read.table("law82.dat")
attach(data)
```

Normaliser les données `GPA` et `LSAT` :

```
GPA=(GPA-mean(GPA))/sd(GPA)
LSAT=(LSAT-mean(LSAT))/sd(LSAT)
```

On considère le modèle de regression

$$X_i = \theta_1 + a_i\theta_2 + \varepsilon_i, 1 \leq i \leq n$$

où a_i, $1 \leq i \leq n$ sont les coordonnées de `GPA` et les variables aléatoires ε_i, $1 \leq i \leq n$ sont i.i.d. de loi Q supposée inconnue. Les paramètres θ_1, θ_2 sont des réels inconnus.

La méthode bootstrap est basée sur l'approximation de la loi $L_n(Q) = \mathcal{L}(\hat{\theta}_2 - \theta_2)$ par $L_n(\hat{Q}_n) = \mathcal{L}(\hat{\theta}_2^\star - \hat{\theta}_2 \,|\, \mathcal{X}_n)$, où $\hat{Q}_n$ est la loi empirique des résidus recentrés.

A. Etude du jeu de données

On suppose que Q est une loi gaussienne. Réaliser une regression de la variable `LSAT` (expliquée) sur la variable `GPA` (explicative) en suivant le protocole suivant :

1) Calculer les estimateurs des moindres carrés $\hat{\theta}$ à l'aide de la fonction `lm`. On pourra également confirmer la valeur de ces estimateurs par un produit matriciel.
2) Tracer `LSAT` en fonction de `GPA` et la droite de regression associée.
3) Calculer les résidus $\hat{\varepsilon}$, puis les résidus recentrés $\hat{e} = \hat{\varepsilon} - \bar{\hat{\varepsilon}}$. Est ce que le modèle de regression gaussien est acceptable ?
4) Proposer un intervalle de confiance de niveau 90% pour θ_2. On pourra utiliser la fonction `confint`. Rejette t-on l'hypothèse nulle H_0: "$\theta_2 = 0$" au niveau 10% ?
5) Proposer un intervalle de confiance de niveau 90% pour θ_2 par bootstrap. Comparer à la question 4).

B. Simulations avec erreurs exponentielles

On suppose maintenant que Q est une loi exponentielle centrée, de paramètre $\lambda > 0$ inconnu (on pourra prendre $\lambda = 2$ dans les simulations). On prend encore a_i, $1 \leq i \leq n$ comme les coordonnées de `GPA` et on choisit $\theta_1 = 3$ et $\theta_2 = 1$. Cependant, nous n'allons pas utiliser les valeurs de `LSAT` dans cette partie.

1) Simuler X suivant le modèle de regression plus haut et répéter les étapes 1)-2)-3) de la partie A. Quelle est la loi des résidus recentrés ?
2) Soit $I(X)$ l'intervalle de confiance en A. 4). Approcher la probabilité de couverture de $I(X)$ à l'aide de 1000 simulations. Que constate t-on ?
3) Construire un intervalle de confiance $I^{boot}(X)$ pour θ_2 de niveau 90% par bootstrap.
Vérifier sur 100 simulations le taux de couverture de $I^{boot}(X)$ (on pourra prendre $B = 100$). Que constate t-on ?