Statistiques non-paramétriques

M2 CEE

Pr. Philippe Polomé, Université Lumière Lyon 2

2016 – 2017

Sommaire

Bootstrap

- Hyp. Bootstrap :
  - Si on pouvait **ré-échantillonner** la pop. dans les mêmes conditions, on obtiendrait un échantillon semblable à celui qu'on a déjà
    - "**Principe de médiocrité**"
    - Pas la même chose que représentativité
- Principe (et 2nde hyp.)
  - Traiter l'éch. comme une pop.
    - Échantillonner l'échantillon de taille n **avec remplacement**
    - S'appelle "Bootstrap par paire" car y et X sont tirés ensemble
  - soit n tirages, chaque i a une probabilité $1/n$ de sortir à chaque tirage
    - On obtient un **échantillon Bootstrap** (de taille n)
    - Certaines obs. sont tirées pls fois, d'autres aucune
    - Hyp. : semblable à ce qu'on aurait obtenu en ré-échantillonnant la pop.

Intervalle de confiance

- Repliquer ce processus B fois
  - B **pseudo-échantillons** différents $< Y_b, X_b >$
- Pour chaque pseudo-échantillon
  - On prend le MRL $Y = X\beta + \epsilon$
  - On calcule un vecteur de valeur estimées $\hat{\beta}_B$
- Si, on prend un élément de β, soit β_i
  - On a B estimations $\hat{\beta}_{ib}$
  - Soit $B = 1000$
    - On ordonne ces 1000 estimations de la plus petite à la plurs grande.
    - Alors les estimations numéro 25 et 975 sont les bornes inf et sup, repsectivement, de l'intervalle de confiance à 95% de β
  - $\widehat{var\left(\hat{\beta}_{ib}\right)} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_{ib} - \bar{\hat{\beta}}_i\right)^2$
    - $\bar{\hat{\beta}}_i = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{ib} = \widehat{E\left(\hat{\beta}_{ib}\right)}$

Pourquoi est-ce intéressant ?

1. Pas d'hypothèse sur la distribution des erreurs
   - 1.1 Mais il ne peut y avoir de corrélation entre observations
   - 1.2 En panels, on ré-échantillonne seulement sur i
     - en utilisant **toutes** les T périodes de chaque i sélectionné
2. On peut calculer des intervales de confiance
   - 2.1 pour toute fonction des paramètres estimés, y-compris non-linéaire
   - 2.2 pour des paramètres estimés de modèles sans propriétés d'échantillons finis connues
     - comme np

- Note : le bootstrap par paires $< Y_b, X_b >$
  - N'est pas la seule façon
    - p.e. on peut se baser sur les résidus
  - Devrait donner des Pairs bootstrap should give reliable standard errors even in the presence of (conditional) heteroskedasticity

Sommaire

Exemple du package AER, *Journals*

- On veut calculer des écarts-types & des intervales de confiance
- data("Journals")
- journals <- Journals[, c("subs", "price")]
- journals$citeprice <- Journals$price/Journals$citations
- jour_lm <- lm(log(subs) ~ log(citeprice), data = journals)

La commande boot( )

- Le bootstrap dans R
  - utilise la commande boot( ) du package boot
  - Elle accepte pls arguments,
    - desquels 3 sont requis :
- Data : les données
- Statistic : une fonction à définir qui renvoie la stat à "bootstrapper"
  - avec 2 arguments
    - Les données (une nouvelle fois !)
  - Un vecteur index qui donne les indices des obs à inclure dans l'échantillon bootstrap
- R : le nombre de réplication
  - B dans la présentation théorique

Construction de la fonction pour l'arg. "statistic" de boot( )

reestim <- function(data, i)

- Se rappeler que dans une ***fonction*** il ne faut qu'énoncer les arguments
  - Ce que la fonction fait est décrit en dessous
- coef(lm(log(subs) ~ log(citeprice), data = data[i,]))
  - reestim est définie pour les besoins de boot( )
    - sur les données et sur un index i des données
  - Ici reestim renvoie les coefficients MCO de log(subs) ~ log(citeprice)
    - et utilise comme index *i* le num. de la ligne des données
    - pas MCO sur la ligne *i*
  - Donc, la fonction boot( ) prend *i* comme index du bootstrap
    - Dans chaque réplication bootstrap, un nouvel éch. extrait des lignes de *data*

Tout ça n'est pas très intuitif,

- mais c'est le format de boot( )

Appeler boot( )

- library("boot")
  - boot est le package recommandé pour le bootstrap
- set.seed(123)
- jour_boot <- boot(journals, reestim, R = 999)
  - boot : 3 arg – données, fonction, B
- jour_boot montre comme résultats :
  - Les coef du lm original
  - La différence avec la moyenne des coeff issus du bootstrap
  - Les écarts-types de ces derniers
    - Desquels on peut calculer les t-stats bootstrappés
- Peu de différence avec la sortie standard coeftest(jour_lm)

Conclusion

- Le bootstrap
  - Très flexible
  - Mais très gourmand en puissance informatique
  - Dans les régressions np que ch suivant, il est la seule technique d'inférence