TP no 3 - Bootstrap

Master parcours SSD - UE Statistique Computationnelle

1 Exercice 1

Dans cet exercice nous allons illustrer la technique du bootstrap sur un cas d'école. Pour cela, commencez par installer le package `bootstrap` pour avoir accès au jeu de données `law` que l'on va utiliser. Ce jeu de données contient deux variables : `LSAT` et `GPA` dont on veut estimer la corrélation. Ces deux variables correspondent à des notes obtenues par des étudiants lors de leur examen d'admission à l'université et tout au long de leurs études secondaires. Une fois chargé le package `bootstrap`, on y accède via `law$GPA` et `law$LSAT`.

1. Calculer le biais et l'erreur type du coefficient de corrélation empirique par une procédure bootstrap de $B = 200$ répétitions.
2. Calculer les intervalles de confiance à 95% obtenus par les deux approches décrites dans le cours (approche "basique" et approche par percentiles)
3. Proposer une représentation graphique des résultats obtenus.
4. Le jeu de données `law` est en fait un sous-échantillon du jeu `law82`. Les intervalles de confiance obtenues sont-ils en accord avec la corrélation estimée à partir du jeu global ?

2 Exercice 2

Reproduire cette analyse en utilisant les fonctions `boot` et `boot.ci` du package `boot`. On rappelle qu'il s'agit ici d'une procédure de bootstrap "ordinaire", et on suggère de porter une attention particulière à l'option `statistic`.

3 Exercice 3

Nous allons travailler sur le jeu de données `cars` qui contient deux variables : la vitesse de voitures des années 1920 et les distance parcourues pour qu'elles s'arrêtent.

1. Visualiser la distance en fonction de la vitesse et superposer le résultat d'une régression linéaire classique (i.e., avec intercept).
   — NB : le jeu de données `cars` fait "nativement" partie de `R`. On accède aux deux variables mentionnées ci-dessus via `cars$speed` et `cars$dist`.
2. Effectuer une régression "bootstrappée" par les approches "par paires" et "par résidus". Considérer $B = 500$ et enregistrer les coefficients obtenus.
   — On pourra en profiter pour visualiser la variabilité obtenue en superposant les droites de régression.
3. Calculer les intervalles de confiance à 95% des coefficients du modèle par la méthode des quantiles.
4. Comparer ces intervalles avec ceux obtenus par l'approche paramétrique classique reposant sur un modèle gaussien. Comment interpréter ces résultats ?
   — On peut (par exemple) obtenir cet intervalle de confiance via la fonction `confint()`.
5. Proposer une représentation graphique permettant de résumer ces résultats.

6. Pour aller plus loin :
   (a) Appliquer une telle procédure bootstrap ("par paires" ou "par résidus", au choix) pour caractériser l'incertitude de prédiction obtenue pour une vitesse égale à 21.
   (b) Comparer la variabilité obtenue par cette procédure bootstrap aux intervalles de confiance proposés par la fonction `predict`, appliquée au modèle de régression linéaire global de la question 1.
      — Pour obtenir ces intervalles de confiance, on utilise la commande suivante :
      ```
      I = predict(global.fit, interval="confidence",
      level=0.95, newdata=data.frame(speed=0:30))
      ```