Bootstrap TP 2

Exercice 1

1. Charger les données `mtcars` et `urine`.
2. Pour les données `urine`, remplacer les valeurs manquantes par la médiane.
3. Calculer les estimateurs classiques des paramètres inconnus.
4. Appliquer les algorithmes “cases sampling“ et “errors sampling“ sur les deux jeux de données.
5. 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)
6. Faire des tests de nullité de chaque paramètre de régression. Comparez les p-values obtenues.

Exercice 2

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. Comparez les tests par boostrap et par appriximation gaussienne pour $\mathcal{H}_0 : \beta_1 = 2$ par Monte Carlo avec $M = 500$ réplications.
3. Changer la loi des ϵ_i pour une loi exponentielle re-centrée de paramètre 2. Cela change-t-il les conclusions ?

Exercice 3

1. Simuler dans un modèle linéaire un échantillon d'apprentissage avec $n = 50$, $p = 5$ avec les mêmes paramètres que dans l'exercice précédent.
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. Faire le choix de modèle par Cp de Mallows (ou AIC, BIC).
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, puis via la cross-validation K-fold (pour K=5, K=2)
6. Choisir un modèle via le bootstrap ($B = 100$) avec des tailles d'échantillons bootstrappés $n' = 25$ et $n' = 50$.
7. Reproduire l'expérience $M = 100$ fois et 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.