INSA – Génie Mathématique Année 2018-2019

Régression Non Linéaire avec Application sous R
GM5 – Feuille TP 2
Bootstrap Non-Paramétrique - Aspects pratiques

L'objet de ce TP est d'expérimenter les méthodes de rééchantillonnage bootstrap.

Avec `R`, on obtient facilement un échantillon bootstrap avec l'option `replace=TRUE` de la commande `sample`. Comparer

```
> sample(1:20,20)
> sample(1:20,20,replace=TRUE)
```

Partie 1 Estimation du biais et de l'erreur standard

On présume, que dans un acier donné, la résistance à la traction, notée R, est liée à la teneur en carbone, notée C. Pour confirmer cette hypothèse, on fait une série de mesures sur des échantillons d'acier. La teneur en carbone est exprimée en $^o/_{ooo}$ et la résistance à la traction de l'acier en kg/mm^2. Les mesures obtenues ont été consignées dans le tableau suivant.

	Echantillons d'acier															
C	72	55	63	38	10	45	77	67	58	74	51	39	27	27	24	32
R	93	81	87	63	38	71	99	91	83	97	76	64	55	56	52	56

Les données sont disponibles à l'URL

`http://lmi2.insa-rouen.fr/~bportier/Data/carbone.txt`

1. Estimer le coefficient de corrélation linéaire.
2. Calculer le biais et l'erreur standard (l'écart-type) du coefficient de corrélation linéaire par la méthode de rééchantillonnage bootstrap par paire (bootstrap des individus). On étudiera la distribution des estimations bootstrap du coefficient de corrélation linéaire en traçant le boxplot et l'histogramme en fréquences. On essaiera plusieurs valeurs du nombre d'échantillons bootstrap. On pourra résumer les résultats dans un tableau.
3. Commenter les résultats obtenus.

Partie 2 Intervalles de confiance pour la prévision

L'observation de la tension systolique T et de l'âge A chez 15 patientes de plus de 40 ans fournit les données suivantes :

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	42	46	71	80	74	70	80	85	72	64	81	41	61	75	53
T	130	125	148	156	162	151	156	162	158	155	160	125	150	165	135

Les données sont disponibles à l'URL

`http://lmi2.insa-rouen.fr/~bportier/Data/systolique.txt`

1. Après une petite étude descriptive des données, proposer un modèle pour ajuster ces données. On argumentera sur le choix du modèle.
2. Avec ce modèle, calculer une prévision de la tension systolique d'une personne agée de 60 ans. A l'aide d'une méthode de rééchantillonnage bootstrap, construire un intervalle de confiance au niveau de confiance 95% pour la prévision. On utilisera la méthode des pourcentiles simples.
3. Même question pour une personne agée de 90 ans.
4. Comparer avec les intervalles fournis par la méthode de Student.

Pour le calcul des quantiles empiriques, on peut utiliser dans `R`, la fonction `quantile`. Essayer par exemple

```
> quantile(rnorm(100000,0,1),c(0.025,0.975))
```

On présentera les résultats dans un seul tableau afin de faciliter la comparaison des résultats obtenus.

Partie 3 Intervalles de confiance

On veut expliquer le maximum de la concentration en Ozone, à l'aide d'un modèle linéaire faisant intervenir 3 variables explicatives : la température à 12h, la nébulosité à 12h et la projection du vent à 12h sur l'axe Est-Ouest,

```
lm(maxO3~  T12 + Ne12  +Vx,data='pollu.txt')
```

Les données sont disponibles à l'URL

`http://lmi2.insa-rouen.fr/~bportier/Data/ozone.txt`

Cet exemple est tiré du livre de P.A. Cornillon & E. Matzner-Løber chez Springer (2007).

1. A l'aide d'une procédure de rééchantillonnage bootstrap, basée sur le bootstrap des résidus, construire des intervalles de confiances au niveau 95% pour les paramètres du modèle. On mettra en œuvre la méthode de l'erreur standard et des pourcentiles simples.
2. Comparer les différents résultats obtenus avec les intervalles de confiance donnés par la fonction `lm`.

Présenter les résultats sous la forme d'un tableau qui permettra une comparaison plus aisée des résultats.