INSA – GM 5 Année 2018-2019

ANTOUN Cyril
ROUFFIAC Jean-Eudes

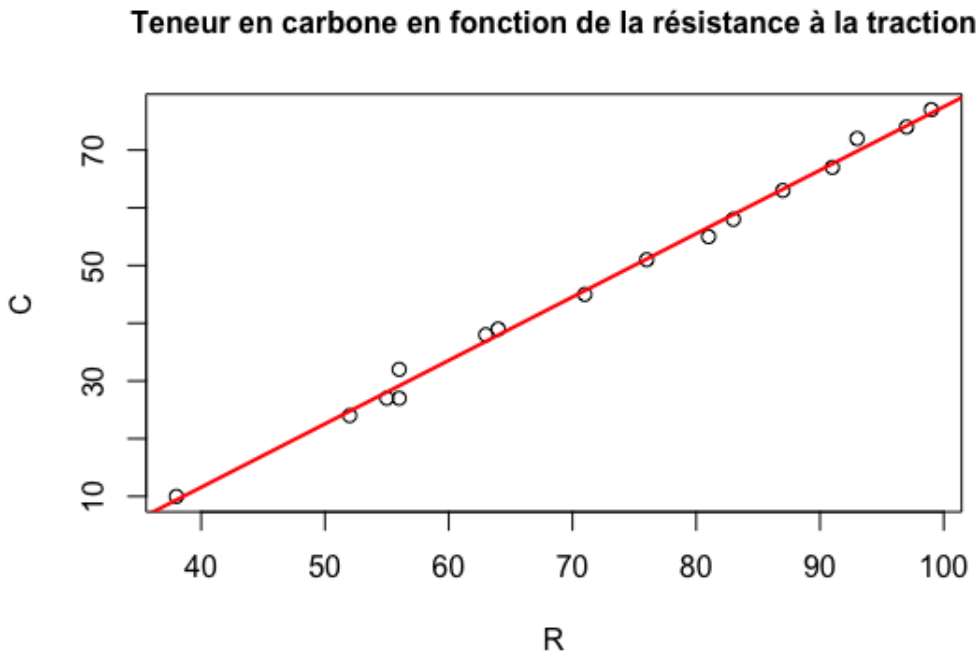
Régression Non Linéaire avec Applications sous R

Rapport de TP n°2

Titre : *Bootstrap Non-Paramétrique - Aspects pratiques*

Partie 1

L'objectif de cette première partie est d'estimer le biais et l'erreur standard à l'aide de la méthode de rééchantillonage bootstrap. Dans cette étude on dispose de mesures de résistance à la traction et de la teneur en carbone relevées sur des échantillons d'acier. On cherche alors à prouver que ces deux variables sont liées. Dans un premier temps on représente la teneur en carbone en fonction de la résistance à la traction sur le graphique suivant :

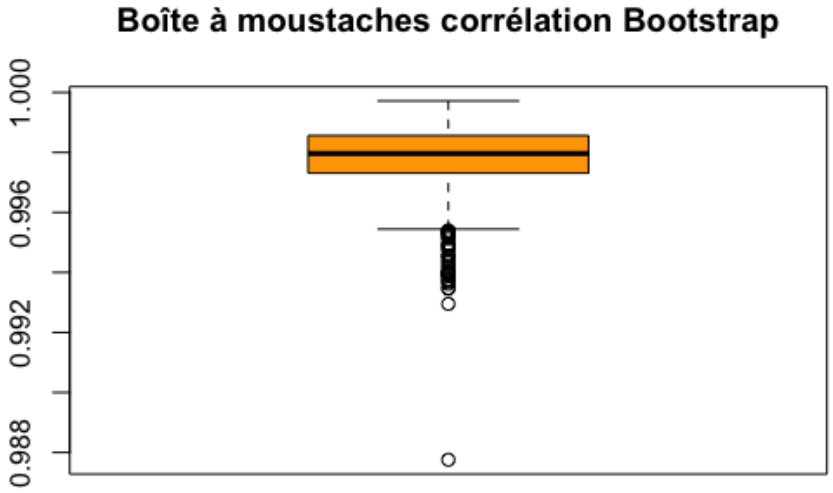


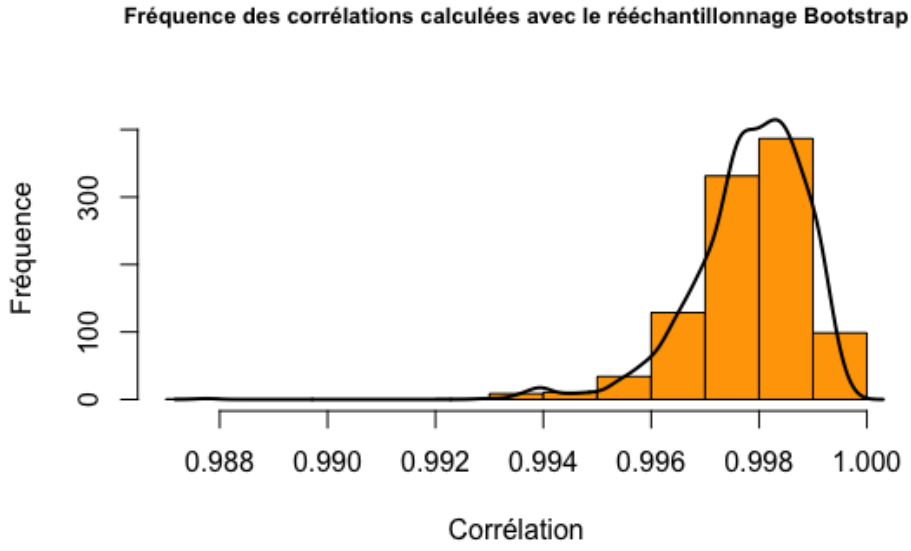
On observe sur ce graphique que le nuage de points forme une droite, l'hypothèse de régression linéaire semble alors se confirmer. De plus, le coefficient de corrélation entre les deux variables vaut 0.9978. Les deux variables sont donc fortement corrélées.

Étant donné que nous avons un faible nombre de données, nous décidons d'utiliser la procédure de rééchantillonage bootstrap par paire afin de savoir si notre estimation du coefficient de corrélation est précise. On calcule donc le biais et l'erreur standard du coefficient de corrélation linéaire pour plusieurs tailles d'échantillons bootstrap. On a résumé les résultats obtenus dans le tableau suivant :

B	50	100	200	500	1000	1500
Biais	$-4.21*10^{-4}$	$-2.74*10^{-4}$	$-6.67*10^{-5}$	$-8.67*10^{-5}$	$-2.63*10^{-5}$	$-3.58*10^{-5}$
Erreur standard	$1.32*10^{-3}$	$1.01*10^{-3}$	$9.65*10^{-4}$	$1.12*10^{-3}$	$1.03*10^{-3}$	$1.03*10^{-3}$

Au vu des faibles valeurs obtenues, on peut en déduire que l'estimation du coefficient de corrélation n'est pas biaisée et que son écart-type, de l'ordre de 10^{-3}, est très faible. Afin de visulaiser la distribution des estimations bootstrap du coefficient de corrélation linéaire, on trace le boxplot et l'histogramme en fréquences. On présente ici les résultats pour une valeur de B de 1500 mais les résultats pour d'autres valeurs de B sont similaires.



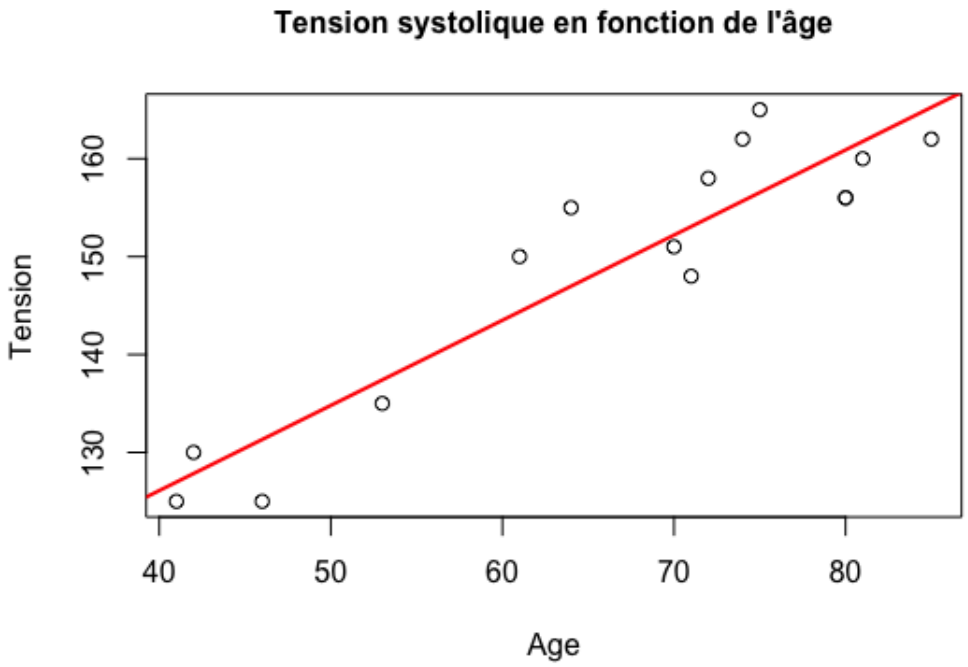


On observe sur ces deux graphiques que la grande majorité des estimations du coeffi-

cient de corrélation se situe entre 0.997 et 0.999. L'hypothèse de forte corrélation est donc confirmée : la résistance à la traction est donc bien liée linéairement à la teneur en carbone pour l'acier étudié.

Partie 2

L'objectif de cette deuxième partie est de connaître l'influence de l'âge sur la tension systolique. Dans cette étude, nous travaillons sur des données relevées sur 15 patients agés de plus de 40 ans. Pour chaque patient i on note y_i sa tension systolique et x_i son âge, avec i allant de 1 à 15. On trace tout d'abord la tension systolique en fonction de l'âge sur le graphique ci-dessous :



La tendance linéaire croissante ainsi que le coefficient de corrélation linéaire de 0.928 laissent penser qu'il peut exister une liaison linéaire entre les deux variables étudiées. On peut alors supposer que l'âge et la tension sont liés par une relation linéaire. On introduit alors un modèle linéaire gaussien, c'est à dire que l'on suppose que les données (y_i) sont les réalisations des 15 variables aléatoires (Y_i) liées aux (x_i) avec $i \in \{1; 2; ...; 15\}$ par la relation :

$$Y_i = \alpha x_i + \mu + \epsilon_i$$

où :

- μ et α_i sont des paramètres réels inconnus,
- les erreurs (ϵ_i) sont des variables aléatoires que l'on suppose indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$

On calcule alors les coefficients estimés du modèle :

- Estimation de α : a = 0.871
- Estimation de μ : b = 91.372

Avec R, on teste l'hypothèse de non-regréssion linéaire et on trouve une p-value de $6.102*10^{-07}$, soit une valeur inférieure à 5%. Par conséquent on rejette au risque 5% l'hypothèse nulle et on déduit que l'âge a une influence significative sur la tension systolique. La part de variance expliquée par le modèle est de 86,14% et celle ajustée de 85,08%. On remarque donc qu'une grande part du modèle est alors expliquée. En conclusion le choix du modèle linéaire gaussien semble approprié.

On peut ainsi estimer la valeur de la tension systolique d'un individu grâce à la relation :

$$y_0 = b + ax_0$$

avec x_0 l'âge d'un patient pour lequel on souhaite estimer sa tension systolique.

Grâce à cette relation, on obtient pour un patient agé de 60 ans une tension systolique de 143.68 mmHg et pour un patient de 90 ans une tension systolique de 169.83 mmHg.

Au vu du faible nombre de données, on peut s'interroger quant à la précision des estimations. On construit alors un intervalle de confiance au risque de 5% pour la prévision à l'aide d'une méthode de rééchantillonnage bootstrap sur les résidus. Dans ce cas on ne fait pas l'hypothèse de normalité pour la loi des erreurs ϵ_i. Ainsi, en appliquant la méthode des pourcentiles simples et pour un B d'une valeur de 1500, on trouve les résultats suivants :

Age	Estimation	Intervalle de confiance
60	143,68	[141.01 ;146.30]
90	169.83	[164.75 ;174.86]

On remarque que les intervalles de confiance ne sont pas très larges comparés aux estimations et semblent précis. On peut considérer au risque 5% que nos estimations sont précises.
Nous souhaitons comparer ces intervalles avec ceux fournis par la méthode de Student. Pour cette méthode les résidus sont supposés indépendants et suivre une loi normale. Nous calculons maintenant les intervalles de confiance à 95% en utilisant la méthode de Student. On obtient les résultats suivants :

Age	Estimation	Intervalle de confiance
60	143,68	[140.43 ;146.93]
90	169.83	[164.05 ;175.61]

On remarque que la méthode de Student donne une largeur d'intervalle sensiblement équivalente à celle donnée par la méthode de rééchantillonage bootstrap.

Partie 3

On cherche dans cette partie à expliquer la variable d'intérêt Y, le maximum de la concentration en Ozone, grâce aux 3 variables explicatives suivantes :

- $x_{1,i}$: La température
- $x_{2,i}$: La nébulosité
- $x_{3,i}$: La projection du vent sur l'axe Est-Ouest

Nous pouvons affirmer après avoir effectuer le test du VIF, que les variables ne sont pas corrélées deux à deux. On souhaite alors expliquer la variable d'intérêt Y par une liaison linéaire à partir des différentes variables explicatives. On introduit donc le modèle linéaire gaussien multiple, c'est à dire que l'on suppose que les données (y_i) sont les réalisations des 91 variables aléatoires (Y_i) liées aux (x_{ji}) avec $j \in \{1; 2; 3\}$ et $i \in \{1; 2; ...; 91\}$ par la relation :

$$Y_i = \mu + \sum_{j=1}^{3} \alpha_j x_{j,i} + \epsilon_i$$

où :

- μ et les α_i sont des paramètres réels inconnus,
- les erreurs (ϵ_i) sont des variables aléatoires que l'on suppose indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$

La p-value du test de non-regression est de $4.433 * 10^{-10}$, ce qui nous permet de dire au risque 5% qu'au moins une des trois variables testées a une influence significative sur le maximum de concentration en ozone.

Notons m l'estimation de μ et a_i les estimations des α_i. Nous avons donc les estimations suivantes : On calcule les estimateurs des α_i et on obtient :

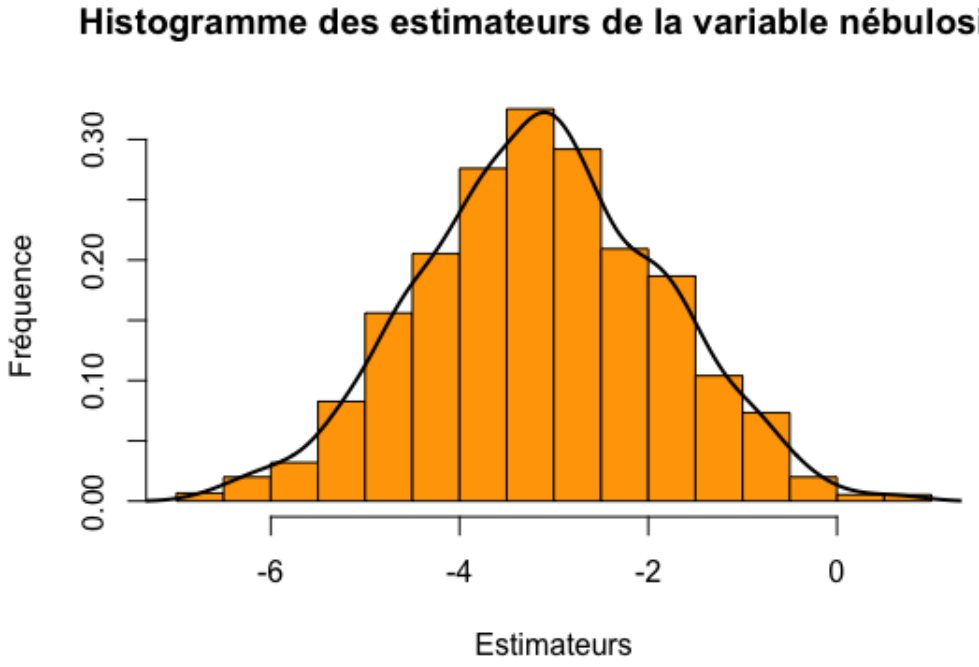
- $m = 37.34$
- $a_1 = 2.16$
- $a_2 = -3.13$
- $a_3 = 3.68$

Nous souhaitons maintenant obtenir un intervalle de confiance au niveau 95% de ces estimateurs. Comme nous ne possédons que très peu de données, nous allons utiliser une procédure de réechantillonnage bootstrap sur les résidus pour un B de 1500. Les deux méthodes de rééchantillonnage bootstrap que nous utilisons sont la méthode de l'erreur standard et celle des pourcentiles simples. Nous comparons également ces résultats avec les intervalles obtenus grâce à la méthode de Student.

Les résultats sont presentés dans le tableau suivant :

Méthode	Erreur standard	Pourcentiles simples	Student
Constante (m)	[-2.66 ;77.34]	[-3.11 ;76.78]	[-4.24 ;78.93]
Température (a_1)	[0.69 ;3.63]	[0.64 ;3.58]	[0.65 ;3.68]
Nébulosité (a_2)	[-5.63 ;-0.64]	[-5.73 ;-0.68]	[-5.75 ;-0.52]
Projection du vent (a_3)	[1.63 ;5.73]	[1.65 ;5.73]	[1.53 ;5.83]

On remarque que les intervalles des méthodes de l'erreur standard et celle des pourcentiles simples sont très proches. Ce résultat était attendu car les distributions des estimateurs des paramètres sont toutes approximativement normales. Comme le montre, par exemple, l'histogramme en fréquences des estimateurs de la variable explicative nébulosité. Les histogrammes des estimateurs des autres paramètres sont semblables. Ils sont représentés en annexe.



La méthode de l'erreur standard s'appuie sur la loi normale. L'allure gaussienne de la répartition des estimations justifie donc la cohérence de trouver des intervalles de confiance équivalents pour la méthode de l'erreur standard et la méthode des pourcentiles simples.

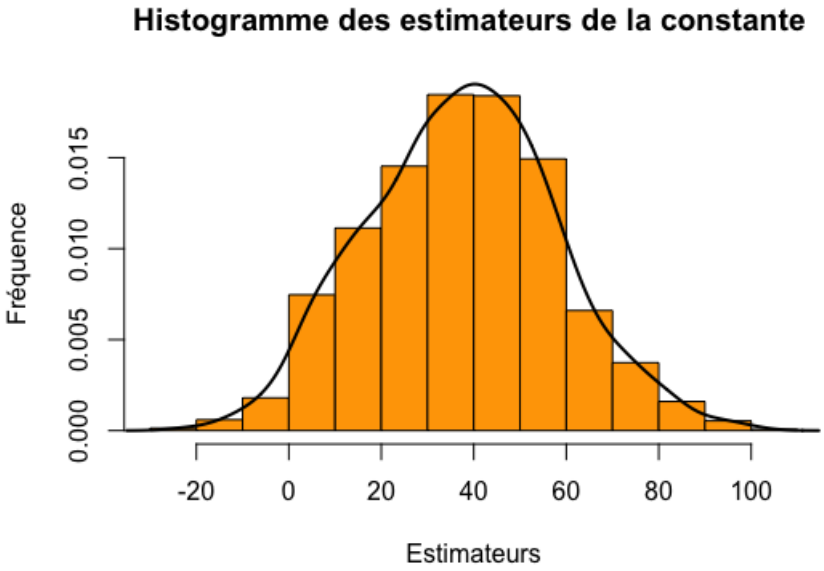
Les intervalles obtenus grâce à la méthode de Student semblent encore une fois proches de ceux déjà obtenus. Nous pouvons vérifier le caractère gaussien des résidus avec un test de Shapiro-Wilk. Ce test nous donne une $p - value$ de 0.1064 qui est supérieure à 5% et qui nous permet d'accepter l'hypothèse de normalité des résidus au risque 5%.

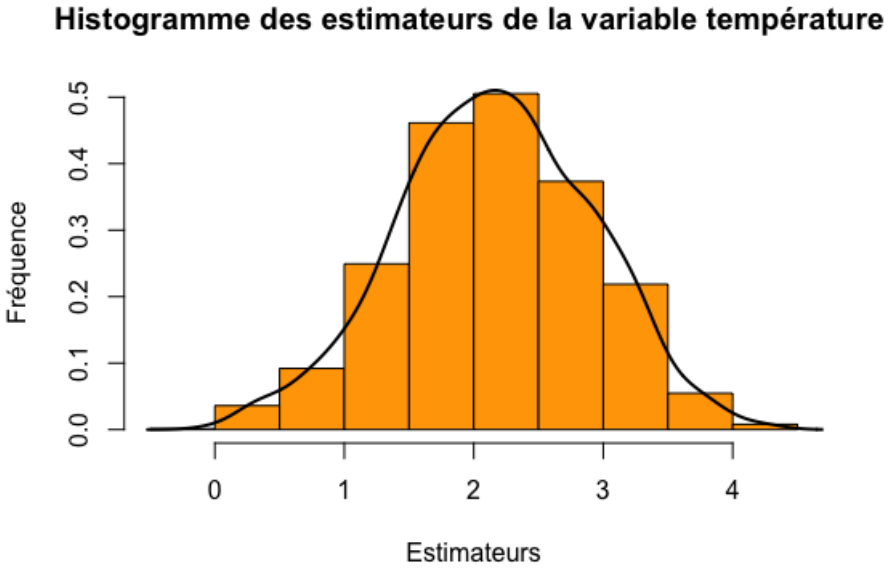
Nous avons donc utilisé toutes nos méthodes sur des résidus qui suivent une loi gaussienne, ce qui justifie la ressemblance entre les trois méthodes.
Si on s'intéresse aux résultats donnés par les méthodes, on remarque que pour chacune d'elles, l'intervalle de la constante m est très large et les intervalles des autres a_i apparaissent peu précis.

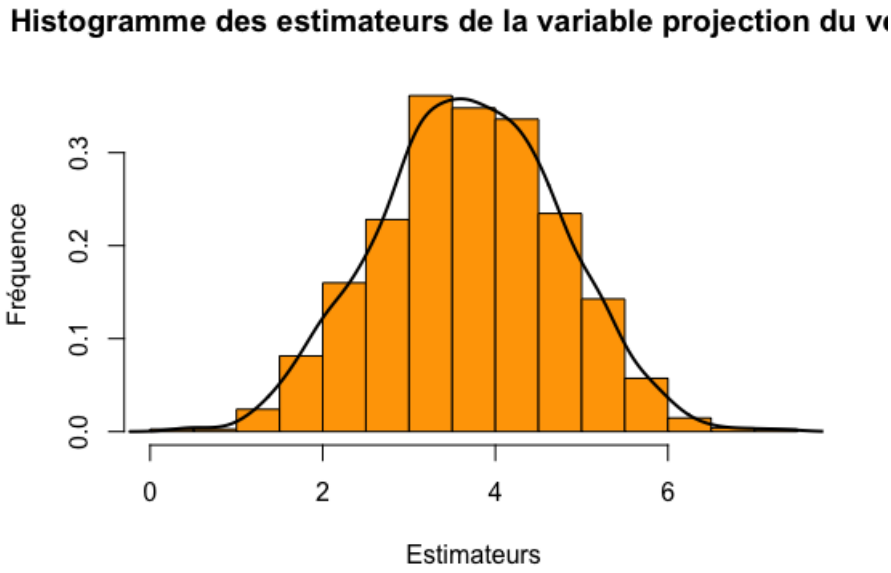
On conclut donc que le modèle linéaire faisant intervenir ces trois variables explicatives n'explique pas bien le maximum de concentration en Ozone, ce qui n'est pas surprenant étant donnée la variance expliquée du modèle de 40%.

Annexe

Les trois autres histogrammes des estimateurs :






Histogramme des estimateurs de la variable projection du v
Fréquence
0.0
0.1
0.2
0.3
0
2
4
6
Estimateurs