Université Paris I, Panthéon - Sorbonne

Deuxième Année Licence M.I.A.S.H.S. 2016 – 2017

TP de Méthodes Numériques n^0 8 : Méthode de Monte-Carlo et régression par moindres carrés

Ce TP a pour but d'introduire la méthode de Monte-Carlo pour approcher une intégrale et d'effectuer des régressions linéaires simples par moindres carrés avec les calculs des estimateurs et tests qui leur sont relatifs.

Approximation d'intégrales par la Méthode de Monte-Carlo

Comme dans le TP8, on désire approcher la valeur d'une intégrale définie:

$$I = \int_a^b f(t)\,dt,$$

où f est une fonction localement intégrable sur $[a,b]$ (c'est-à-dire que f est définie sur tout $[a',b']$, où $a < a' < b' < b$ et $\int_{a'}^{b'} f(t)\,dt$ existe). Alors

$$I = (b-a)\,\mathbb{E}(X) \qquad \text{où} \qquad X \sim \mathcal{U}([a,b]) \quad \text{(à vérifier!)}.$$

Grâce à la loi des grands nombres et au théorème de la limite centrale, il est facile de voir que si on note:

$$\widehat{I}_n = \frac{(b-a)}{n}\left(X_1 + \cdots + X_n\right) \qquad \text{où les } (X_i) \text{ sont des v.a.i.i.d. de loi U([a,b])}$$

alors $\widehat{I}_n \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} I$, $\sqrt{n}(\widehat{I}_n - I) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,\sigma^2)$, où $\sigma^2 = (b-a)\int_a^b f^2(t)\,dt - I^2$ (à vérifier). Ainsi, comme le quantile à 97.5% d'une loi gaussienne est d'environ 1.96, alors un intervalle de confiance à 95% sur I est:

$$\left[\widehat{I}_n - 1.96\,\frac{\widehat{\sigma}_n}{\sqrt{n}}\,,\, \widehat{I}_n - 1.96\,\frac{\widehat{\sigma}_n}{\sqrt{n}}\right] \qquad \text{où} \quad \widehat{\sigma}_n^2 = \frac{(b-a)^2}{n}\sum_{i=1}^n f^2(X) - \widehat{I}_n^2.$$

Expliquer cette intervalle de confiance.
On applique cette méthode à l'estimation de $I = \frac{1}{\sqrt{2\pi}}\int_0^1 \exp(-t^2/2)\,dt$. Ecrire le programme permettant d'obtenir l'intervalle de confiance précédent pour $n = 1000$, puis $n = 10^6$. Que pensez-vous de la qualité de l'estimation obtenue? Quelle est l'ordre de grandeur de l'erreur en fonction de n? Comparer avec les méthodes numériques utilisées lors du TP8...

L'utilisation de la méthode de Monté-Carlo pour estimer la valeur numérique d'une intégrale n'est intéressante e"'n réalité que dans 2 cas: lorsque la fonction f est très irrégulière, par exemple discontinue ou avec des pôles, ou bien lorsque l'on veut approcher une intégrale multiple d'une fonction pas trop régulière. Sinon, des méthodes numériques déterministes comme celles de Simpson, ou d'autres encore plus performantes, sont à conseiller.

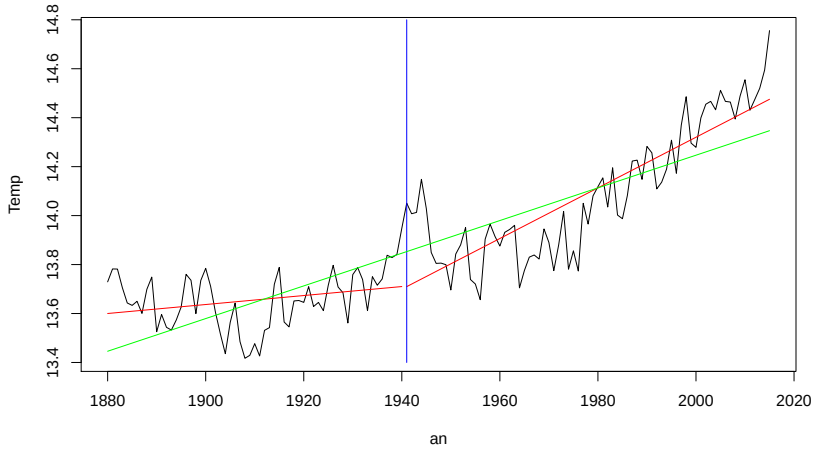
Voici un exemple où utiliser la méthode de Monte-Carlo peut avoir du sens:

$$I = \int_0^1 |t - \ln(2)|^{0.1}\,dt.$$

Calculer théoriquement la valeur de I, puis approcher cette valeur avec les 3 méthodes numériques déterministes, pour différentes valeurs de n. Retrouve-t-on que la méthode de Simpson donne de meilleurs résultats? Comaprer ensuite avec des estimations données par la méthode de Monte-Carlo. Qu'en pensez-vous? L'intervalle de confiance obtenu précédemment est-il encore possible?

Régression linéaire simple par moindres carrés

On se place dans le cadre où l'on a observé (X_i, Y_i) pour $i = 1, \cdots, n$, où X et Y sont deux variables. On suppose que globalement les Y_i se comporte linéairement en fonction des X_i, c'est-à-dire que stricto-sensu les (X_i, Y_i) ne sont pas tous sur la même droite, mais ils n'en sont pas très éloignés.
On peut prendre comme exemple un cas particulier tout à fait d'actualité et concernant le réchauffement climatique: on considère les températures annuelles moyennes du globe depuis 1880:



Sur le graphe précédent on a tracé en noir l'évolution des températures chaque année, et en vert on voit une droite qui semble assez bien résumer l'évolution de ces températures (les deux autres droites rouges donnant l'évolution entre 1880 et 1940, et entre 1941 et 2015). La question posée est alors celle-ci: comment déterminer une telle droite, c'est-à-dire une droite qui semble coller au mieux aux données?
Supposons donc que l'on considère une droite qui s'écrit $Y = a\,X + b$ et qu'on la sur-rajoute au nuage des points (X_i, Y_i). On peut alors définir une distance au carré entre les points et la droite de la manière suivante:

$$d_n^2(a,b) = \sum_{i=1}^{n} \big(Y_i - (a\,X_i + b)\big)^2.$$

D'autres distances seraient possibles, mais celle-ci possède certaines propriétés qui la rende plus intéressantes que d'autres (en particulier la possibilité d'avoir des estimateurs et tests explicites). Il est clair que pour rendre la droite la plus proche possible des données, on a envie de minimiser cette distance carrée et on définit ainsi:

$$(\widehat{a}_n\,,\,\widehat{b}_n) = \underset{(a,b)\in\mathbf{R}^2}{\text{Argmin}}\; d_n^2(a,b).$$

On appelle ainsi $\widehat{a}_n$ et $\widehat{b}_n$ les estimateurs de a et b par **moindres carrés**. On peut alors montrer (le faire!) que:

$$\widehat{a}_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}_n\,\overline{Y}_n}{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X}_n)^2}, \qquad \widehat{b}_n = \overline{Y}_n - \widehat{a}_n\,\overline{X}_n, \quad \text{avec} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{et} \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i.$$

Pour jauger de la qualité d'ajustement de la droite de régression on utulise souvent un coefficient, appelé **coefficient de détermination** R^2 défini par:

$$R^2 = 1 - \frac{\sum_{i=1}^n \big(Y_i - (\widehat{a}_n\,X_i + \widehat{b}_n)\big)^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}.$$

Il est possible de montrer que l'on a toujours $0 \leq R^2 \leq 1$ et il est bien clair que plus R^2 est proche de 1, plus les données "collent" à la droite de régression. Cependant, à moins que $R^2 = 1$, on ne peut en général rien conclure quant à la légitimité d'une représentation des données par une droite à partir de la valeur prise par R^2. En revanche, on peut montrer le résultat suivant:

Test de Student: Si $\frac{\max_{1\leq i\leq n}(\overline{X}_n - X_i)^2}{\sum_{i=1}^n X_i^2 - \overline{X}_n^2} \underset{n\to+\infty}{\longrightarrow} 0$, alors $\widehat{T}_n = \widehat{a}_n \sqrt{\frac{\sum_{i=1}^n X_i^2 - (\overline{X}_n)^2}{\sum_{i=1}^n \big(Y_i - (\widehat{a}_n\,X_i + \widehat{b}_n)\big)^2}} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0\,,\,1)$.
Ainsi, quand n est grand, si $|\widehat{T}_n| > 1.96$, la modélisation par une droite est légitime. Sous cette hypothèse, on acceptera le fait que si x est une nouvelle valeur de X, alors la valeur prédite de Y sera $\widehat{y} = \widehat{a}_n\,x + \widehat{b}_n$.

Pour illustrer cette méthode et ce dernier résultat, on effectue une simulation:

```
n=200; X=c(1:n); erreur=rnorm(n,0,64);
Y=-5+2*X+erreur
plot(X,Y)
a=(mean(X*Y)-mean(X)*mean(Y))/(mean(X^2)-mean(X)^2);  b=mean(Y)-a*mean(X)
lines(X,a*X+b,col='red')
R2=1-sum((Y-a*X-b)^2)/sum((Y-mean(Y))^2)
T=a/sqrt(1-R2)
a; b; R2; T
```

Vérifier que a n'est pas trop éloigné de la valeur attendue (laquelle?). Qu'en est-il pour b? Que conclure de la valeur de T? Est-ce logique? Pour améliorer la pertinence des résultats obtenus, relancer 100 fois le programme précédent en mémorisant les différentes statistiques. Tracer ensuite des histogrammes. Qu'en pensez-vous? Prédire la valeur de Y lorsque $X = n + 20$. Recommencez tout cela en changeant la valeur de n (par exemple avec $n = 10$ et $n = 1000$). Qu'en pensez-vous? Dans le cas de cette simulation, avait-on bien la condition suffisante permettant la convergence de $\widehat{T}$?

Enfin, on travaille sur les données de températures représentées un peu plus haut. Pour obtenir ces données, aller sur `https://samm.univ-paris1.fr/L2-MIASHS-Methodes-Numeriques-S4`, et télécharger sur votre ordinateur dans un répertoire `C:|Donnees` le fichier de données de températures `TGlobe1.txt`. Pour intégrer directement ces données sous R, il vous faut taper les commandes:

```
TG=read.table('C:/Donnees/TGlobe1.txt',header=TRUE,sep = "\t")
TG; names(TG); TG$An; TG$Temp
An=as.numeric(TG$An); Temp=as.numeric(TG$Temp)
```

Ainsi TG est d'abord crée dans le format R appelé `data.frame` et a deux variables associées An et $Temp$. La commande `as.numeric` permet de convertir ces variables initialement considérées comme qualitatives en des variables quantitatives. Analyser ces données pour obtenir la figure affichée plus haut. Les tests de Student sont-il vérifiés pour les 3 régressions effectuées? Donner une prédiction pour la température en 2100 avec les deux modèles de régression considérés (celui sur la période allant 1880 à 2015, l'autre allant de 1941 à 2015).

Exercices

1. Avec la méthode de Monte-Carlo, calculer $\int_{-1}^{1}(1-t^2)^{-1/2}dt$ et en déduire une approximation de π. Est-ce possible d'obtenir une approximation à 10^{-15} près?
2. Reprendre la simulation réalisée plus haut, mais cette fois-ci en considérant la commande `erreur=runif(n,-15,15)`.
3. Reprendre la simulation réalisée plus haut, mais cette fois-ci en considérant la commande `X=runif(n,-3,15)`.
4. Reprendre la simulation réalisée plus haut, mais cette fois-ci avec la commande `X=1/c(1:n)`.