Université Paris Dauphine — Année 2012-2013
Département MIDO
L3 - Statistique Mathématique

Projet de statistique mathématique. Troisième partie.

Le projet compte pour 30% de la note finale de Statistique Mathématique.

Déroulement du projet : La troisième partie du projet doit être déposée au bureau B530 aux horaires d'ouvertures avant le 2 mai. Tout retard équivaut à un 0 pour cette partie du projet. Cette troisième partie compte pour 8/20 de la note finale du projet.

Les étudiants doivent utiliser le logiciel R [1]. Les principales fonctions utilisées doivent apparaître dans le rapport.

Dans la première partie du projet, chaque binôme ou trinôme a traité une série de données réelles provenant des volumes au bid dans un carnet d'ordre. Les données ainsi obtenues $x_1, \ldots, x_n$ sont les réalisations d'observations $X_1, \ldots, X_n$ supposées iid. Dans la seconde partie du projet, nous avons estimer analytiquement et numériquement les paramètres inconnus pour le modèle GEV. Les estimateurs par la méthode des moments et par la méthode du maximum de vraisemblance ont été comparés graphiquement par rapport à l'histogramme.

1 Objectif de la troisième partie

Dans cette troisième partie, des procédures de test sont mises en place pour décider quel estimateur choisir, si la modélisation est adéquate au données où non et si la modélisation proposée dans la littérature est plus adéquate (pour votre jeu de données) que celle que vous avez choisie.

Le test utilisé est celui du χ^2. Nous allons distinguer trois types de procédures :

- Pour chaque estimation fournie dans la deuxième partie du projet, nous allons construire un test d'adéquation du χ^2 de la loi correspondante. L'estimation la plus favorable est déterminée en comparant les $p-$valeurs obtenues pour les différents tests,
- Tester l'adéquation du modèle choisi dans la première partie du projet (avec la meilleur estimation d'après le point précédent) à partir d'un tests d'adéquation du χ^2 du modèle,
- Comparer la p-valeur du test précédent avec la $p-$valeur d'un test d'adéquation du modèle utilisé dans la littérature.

1. voir ci-besoin introduction à R : http ://www.ceremade.dauphine.fr/%7Exian/Noise/R.pdf ou le site officiel http ://www.r-project.org/

2 Comparaison d'estimations par les p-valeurs

Le but de cette section est de mettre en place des tests d'adéquation pour chacune des lois correspondantes aux différentes estimations de paramètres obtenues dans la deuxième partie du projet. La comparaison des p-valeurs de ces tests permet de déterminer quelle est la meilleure estimation de la deuxième partie. Vous devez donc au maximum réaliser 4 tests d'adéquation du χ^2 différents.

Le test d'adéquation du χ^2 correspond à la problématique (pour plus de détails on réfère au chap. 11 du polycopié "Statistique mathématique")

$$H_0 : \quad \mathbf{p} = \mathbf{q} \qquad \text{contre} \qquad H_1 : \quad \mathbf{p} \neq \mathbf{q},$$

où $\mathbf{q}$ est le vecteur des proportions correspondant au modèle à N classes issu de la loi $P_{\hat{\theta}_n}$ pour la loi paramétrique P_θ obtenue dans la partie 1 du projet et une des 4 estimations $\hat{\theta}_n$ obtenue dans la partie 2 du projet.

Question 1 : On se fixe le nombre de classes à $N = 10$. Calculer la fréquence empirique pour une partition de $\mathbb{R}$ en 10 intervalles $I_1, \ldots, I_{10}$ en utilisant la fonction "hist" sous R. On obtient ainsi un vecteur

$$\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_6) = \left(\frac{1}{n} \sum_{i=1}^{n} 1_{x_i \in I_1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \in I_6} \right).$$

Attention : vérifier que les intervalles sont tels que $\hat{p}_i > 5\%$ pour chaque i, i.e. tels que les données x_i soient suffisamment présentes dans chacun d'entre eux.

Question 2 : A l'aide de la méthode de Monte-Carlo sur $M = 10000$ simulations, donner une approximation numérique des vecteurs $\mathbf{q} = (P_{\hat{\theta}_n}(I_1), \ldots, P_{\hat{\theta}_n}(I_{10}))$ pour chaque estimations $\hat{\theta}_n$ obtenues dans la partie 2. En déduire la valeur des statistiques du χ^2 pour les 4 tests d'adéquation.

Question 3 : Calculer les p-valeurs associées à ces tests. En comparant ces $p-$valeurs, déduire quelle est la meilleure estimation $\hat{\theta}_n$ obtenue dans la partie 2 du projet. Cette conclusion est-elle conforme à celle faite graphiquement dans la deuxième partie du projet ?

On rappelle que, pour un test d'adéquation du χ^2 à une loi, sa p-valeur est égale à $1 - F(\hat{\chi}_n^2)$, $\hat{\chi}_n^2$ étant la statistique du χ^2 associée au test et F étant la fonction de répartition d'une χ^2_{N-1}.

Remarque : si les p-valeurs obtenues sont trop faibles pour être distinctes de l'erreur d'arrondi de la machine, préférer utiliser un argument de monotonicité et comparer directement les valeurs des statistiques du χ^2 entre elles.

3 Adéquation du modèle choisi

Nous allons tester si le modèle choisi dans la partie 1 du projet est en adéquation avec les données observées. On utilise pour cela le test d'adéquation du χ^2 d'un modèle

$$H_0 : \quad \exists \theta \in \Theta \ / \ \mathbf{p} = \mathbf{p}(\theta) \qquad \text{contre} \qquad H_1 : \quad \forall \theta \in \Theta \ / \ \mathbf{p} \neq \mathbf{p}(\theta)$$

où $\mathbf{p}(\theta)$ est le vecteur des proportions correspondant au modèle à N classes issues du modèle GEV (P_θ, Θ) associé à $\theta = \hat{\theta}_n$ où $\hat{\theta}_n$ est l'estimation retenue dans la section précédente (i.e. celle dont la p-valeur est la plus grande).

Question 4 On garde le même nombre de classes $N = 10$ ainsi que la même partition que dans la Question 1. Montrer que la statistique du χ^2 pour ce test a déjà été calculée. Calculer la p-valeur de ce test d'adéquation du modèle. Le modèle GEV vous semble-t-il adéquat ?

4 Comparaison avec un modèle utilisé dans la littérature

Comme énoncé dans la partie 1 du projet, la distribution du carnet d'ordre a un impact sur la stratégie d'achats ou de ventes d'actifs. En mathématiques financières, différentes hypothèses simplificatrices ont été faites sur cette distribution afin de pouvoir calculer une stratégie optimale d'achats ou de ventes. Vous allez comparer le modèle GEV avec le modèle gamma (densité 6 dans le tableau des densités) qui a été retenu dans l'article

"Statistical properties of stock order books : empirical results and models", de Bouchaud, Mézard et Potters.

Question 5 Calculer la statistique du χ^2 d'adéquation du modèle utilisé dans la littérature avec les mêmes classes que précédemment (cela implique l'estimation de paramètres). Calculer la $p-$valeur et la comparer avec la $p-$valeur calculée dans la section précédente. Peut-on comparer directement les statistiques de tests du χ^2 comme dans le cas du test d'adéquation à une loi (c.f. remarque précédente) ?