1 Exercices

Exercice 1 (Modèle de translation et d'échelle). Cet exercice est élémentaire. Son unique objectif est de vous faire jouer entre différentes façons de spécifier les modèles statistiques. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et g une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On note $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

1. Soit $(\zeta_1, \ldots, \zeta_n)$ n variables aléatoires réelles, définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, indépendantes et de même loi de densité g par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Montrer que pour tout $(\mu, \sigma) \in \Theta$, la loi du vecteur aléatoire

$$(\mu + \sigma\zeta_1, \ldots, \mu + \sigma\zeta_n)$$

a une densité $p_{n,\theta}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ que l'on déterminera.

Il est d'usage d'appeler μ le paramètre de *translation* et σ le paramètre d'*échelle*.

2. On considère le modèle statistique

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{ p_{n,\theta} \cdot \lambda_{\text{Leb}}^{\otimes n} \,:\, \theta \in \Theta \right\}\right)$$

On note $(X_1, \ldots, X_n)$ les observations : pour tout $i \in \{1, \ldots, n\}$, $X_i(x_1, \ldots, x_n) = x_i$ où $(x_1, \cdots, x_n)$ désignent les données collectées. Montrer que les statistiques $(X_1, \ldots, X_n)$ sont indépendantes et identiquement distribuées.

3. Montrer que pour tout $\theta = (\mu, \sigma) \in \Theta$, les variables aléatoires réelles

$$\frac{X_i - \mu}{\sigma}, \quad i \in \{1, \ldots, n\}$$

sont, sous $p_{n,\theta} \cdot \lambda_{\text{Leb}}^{\otimes n}$, indépendantes et identiquement distribuées de loi de densité g par rapport à la mesure de Lebesgue.

4. Supposons que g est une densité gaussienne centrée réduite. Déterminer le modèle statistique induit par les statistiques $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$. Proposer un estimateur de μ et σ^2.

Exercice 2. Nous cherchons à modéliser sur une population de n individus la dépendance d'une réponse par rapport à des variables explicatives. On collecte donc

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \,,$$

où $y_i \in \mathbb{R}$ et $\mathbf{x}_i \in \mathbb{R}^k$ sont respectivement la *réponse* et les *variables explicatives* pour le i-ème individu.

Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité, g une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et $f : \mathbb{R} \to \mathbb{R}$ une fonction. Posons $\theta = (\beta, \sigma) \in \Theta = \mathbb{R}^k \times \mathbb{R}_+^*$.

1. Soit $(\xi_1, \ldots, \xi_n)$ n variables aléatoires réelles définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, indépendantes et de loi de densité g. Montrer que pour tout $\theta \in \Theta$, la loi du vecteur aléatoire

$$\left(f(\beta' \mathbf{x}_1) + \sigma\xi_1, \ldots, f(\beta' \mathbf{x}_n) + \sigma\xi_n\right)$$

a une densité $p_{n,\theta}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, que l'on déterminera.

2. Considérons le modèle statistique suivant

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{ p_{n,\theta} \cdot \lambda_{\text{Leb}}^{\otimes n} \,:\, \theta = (\beta, \sigma) \in \Theta = \mathbb{R}^k \times \mathbb{R}_+^* \right\}\right).$$

Pour $i \in \{1,\dots,n\}$ nous notons Y_i la i-ème variable canonique, $Y_i(y_1,\dots,y_n) = y_i$. Montrer que les statistiques $(Y_1,\dots,Y_n)$ sont indépendantes et préciser leur loi.

3. Montrer que pour tout $\theta \in \Theta$, les variables

$$\sigma^{-1}\{Y_i - f(\beta' \mathbf{x}_i)\}, \qquad i \in \{1,\cdots,n\}$$

sont, sous $p_{n,\theta} \cdot \lambda_{\text{Leb}}^{\otimes n}$, indépendantes de loi de densité g par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

4. Dans cette question, $k = 2$, g est la densité d'une loi gaussienne centrée réduite, $f(\eta) = \eta$ et pour tout $i \in \{1,\cdots,n\}$

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix} \qquad f(\beta' \mathbf{x}_i) = \beta_0 + \beta_1 x_i.$$

Proposer un estimateur de β_0 et β_1.

5. Nous disposons de 150 mesures de concentration d'Ozone (moyenne observées entre 13 :00 et 15 :00 à Roosevelt Island en p.p.m) en fonctions de différents facteurs : radiations solaire, vitesse du vent, température. Nous considérons un modèle à un facteur : la réponse concentration d'Ozone et la variable explicative est la radiation solaire. Le modèle vous

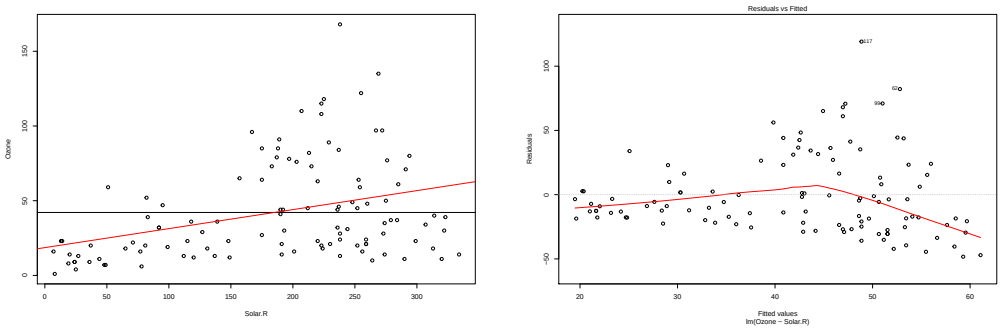

FIGURE 1 – figure de gauche : les données et la droite régression. Figure de droite : les résidus de régression

semble-t-il approprié ? Quelles améliorations du modèle vous semblerez souhaitable ?

Exercice 3 (Score de football). On appelle *loi de Poisson* de paramètre $\lambda > 0$ la loi de densité $(p(\lambda,k), k \in \mathbb{N})$ par rapport à la mesure de comptage μ sur $\mathbb{N}$:

$$p_\lambda(k) = p(\lambda,k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}$$

1. Calculer la fonction génératrice des moments de la loi de Poisson de paramètre λ.
2. En déduire la moyenne et la variance d'une loi de Poisson de paramètre $\lambda \in \mathbb{R}^+$.
3. Montrer que si X_1 et X_2 sont deux variables indépendantes de loi de Poisson de paramètres $\lambda_1 > 0$ et $\lambda_2 > 0$, alors la variable aléatoire $X_1 + X_2$ suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle

$$(\mathbb{N}, \mathscr{P}(\mathbb{N}), \{p_\lambda \cdot \mu \,:\, \lambda \in \mathbb{R}_+^*\})$$

4. Définir le modèle statistique induit par la statistique $\sum_{i=1}^n X_i$.
5. Proposer une méthode d'estimation du paramètre λ.

On se propose de modéliser le nombre de buts inscrits dans un match de football par une loi de Poisson. On considère tout d'abord que le nombre de buts inscrits par l'équipe locale et l'équipe visiteuse sont deux variables de Poisson indépendantes de loi de Poisson de paramètres différents. On suppose aussi que les résultats des matchs sont indépendants.

6. Construire le modèle statistique associé à l'observation des résultats de n matchs.
7. On a collecté les buts marqués en *premier league* de la saison 2004-2005 à la saison 2008-2009. Le nombre de matchs est de $n = 1900$:

   le nombre de buts marqués est en moyenne 2.523 avec une variance de 2.640,

   le nombre de buts marqués par l'équipe locale est en moyenne 1.468 buts avec une variance de 1.617,

   le nombre de buts marqués par l'équipe visiteuse est en moyenne 1.055 avec une variance de 1.158.

   Proposer un estimateur des intensités λ_1 et λ_2 des deux processus de Poisson introduits pour modéliser le nombre de buts marqués par chacune des deux équipes. Pourquoi l'hypothèse poissonienne est-elle discutable ?

Au lieu d'ajuster une loi de Poisson, il semble plus judicieux dans ce cas de considérer une famille de loi présentant une "sur-dispersion" par rapport à la loi de Poisson (i.e. pour laquelle la variance puisse être plus grande que la moyenne). Etudions plus en détail le nombre de buts marqués par l'équipe locale et l'équipe visiteuse (voir figure 2).

L'analyse de ces résultats suggère que la modélisation poissonnienne sous-estime le nombre de scores nuls et sur-estime en contre-partie les cas où 1 ou 2 buts sont marqués. On considère comme modèle, un mélange de la distribution de Poisson et d'un atome en 0,

$$p_{\pi,\lambda}(k) = (1-\pi)\mathbb{1}_{\{0\}}(k) + \pi \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,,$$

où $\pi \in]0,1[$ est la proportion du mélange.

8. Calculer la moyenne et le moment d'ordre 2 de cette distribution.
9. Proposer une méthode d'estimation de π et de λ.

	Observés	Poisson ($\lambda = 1.468$)
0	469	437.7
1	621	642.6
2	456	471.7
3	217	230.8
4	100	84.7
≥ 5	37	32.5
Total	1900	

	Observés	Poisson ($\lambda = 1.055$)
0	692	661.6
1	680	697.9
2	335	368.2
3	131	129.5
4	51	34.1
≥ 5	11	8.7
Total	1900	

FIGURE 2 – Résultats de l'équipe locale (gauche) et de l'équipe visiteuse (droite)

Exercice 4 (Prix d'un actif financier). Soit X une variable aléatoire réelle d'espérance μ et d'écart type σ. Si $\mathbb{E}[|X|^3] < \infty$, on définit le coefficient d'asymétrie comme le moment d'ordre trois de la variable centrée réduite :

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\mu_2^{3/2}} ,$$

avec μ_i les moments centrés d'ordre i. Si $\mathbb{E}[X^4] < \infty$, on définit son kurtosis non normalisé (coefficient d'aplatissement) comme le moment d'ordre quatre de la variable centrée réduite :

$$\beta_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\mu_2{}^2} .$$

On définit l'excès de kurtosis comme $\gamma_2 = \beta_2 - 3$.

On observe la suite $p_1, \cdots, p_n$ du prix d'un actif financier à la clôture d'un marché (prix journalier, voir Figure 3). On modélise ces données comme une réalisation du vecteur aléatoire $(P_1, \cdots, P_n)$. On appelle log-rendement de cet actif la quantité

$$X_k = \log(P_k/P_{k-1}) ;$$

voir Figure 4. Un modèle couramment utilisé (associé à la théorie proposée par F. Black, M. Scholes et R. Merton, prix Nobel 1997) consiste à supposer que

(i) Les log-rendements $\{X_i, i \geq 1\}$ sont indépendants et identiquement distribués.

(ii) et ils suivent une distribution gaussienne de moyenne μ et de variance σ^2 inconnues.

1. Proposer un modèle statistique des log-rendements.
2. Proposer un estimateur de la moyenne μ et de la variance σ^2.
3. On superpose l'histogramme des observations et la densité gaussienne estimée (voir figure 5). Peut-on être satisfait de ce modèle ? Qu'observe-t-on ?
4. Lorsque X est une variable aléatoire gaussienne de moyenne μ et de variance σ^2,
   a) montrer que le coefficient d'asymétrie est nul.

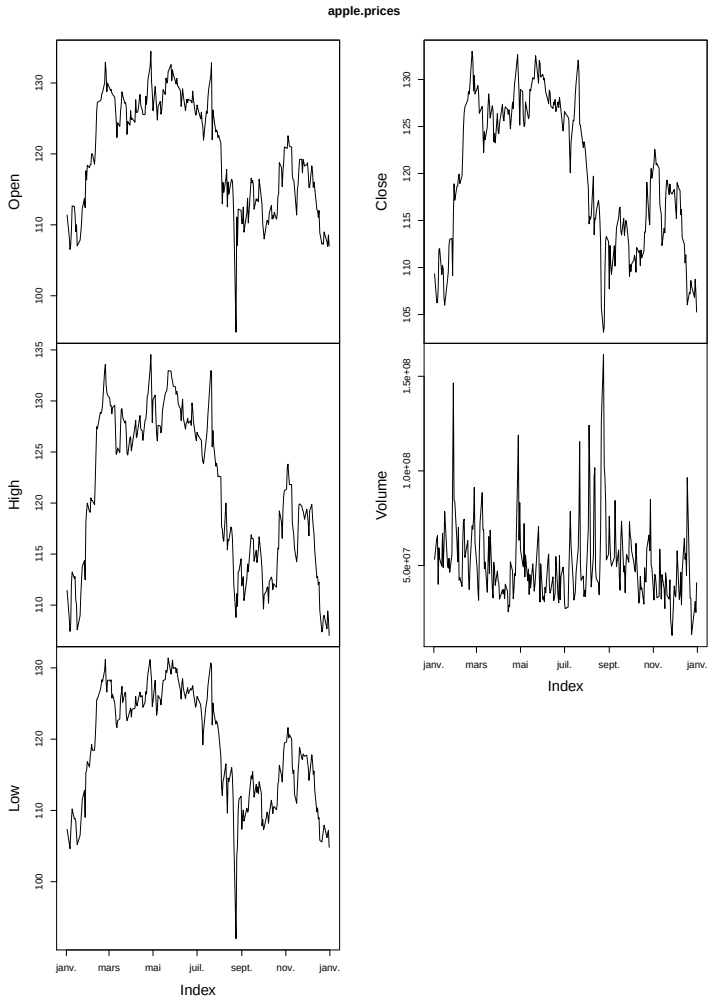

FIGURE 3 – Exercice 4 - Prix

b) calculer $\mathbb{E}[e^{tX}]$ pour tout $t \in \mathbb{R}$. En identifiant les premiers termes du développement de $\log \mathbb{E}[e^{tX}]$, montrer que l'excès de kurtosis est nul.

5. Proposer un estimateur empirique du coefficient d'asymétrie et du coefficient d'excès de kurtosis.
6. En évaluant ces estimateurs sur la série des log-rendements, nous obtenons -0.0707 pour l'asymétrie et 1.274035 pour l'excès de kurtosis. Le modèle retenu vous semble-t-il acceptable ?
7. Pour modéliser les log-rendements, R. Engle (Prix Nobel d'Economie 2003) a proposé le modèle ARCH(1) :

$$X_k = \sqrt{\alpha_0 + \alpha_1 X_{k-1}^2} Z_k \,, \quad X_0 = 0 \,,$$

où $\alpha_0 > 0$ et $\alpha_1 \geq 0$ et $\{Z_k\}_{k=1}^n$ est une suite i.i.d. de variables aléatoires gaussiennes

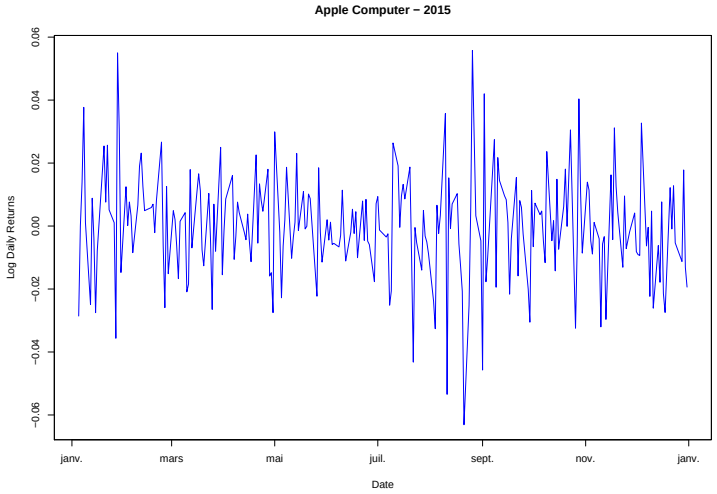

FIGURE 4 – Exercice 4 - Log-rendements

centrées réduites. Définir le modèle statistique associé.

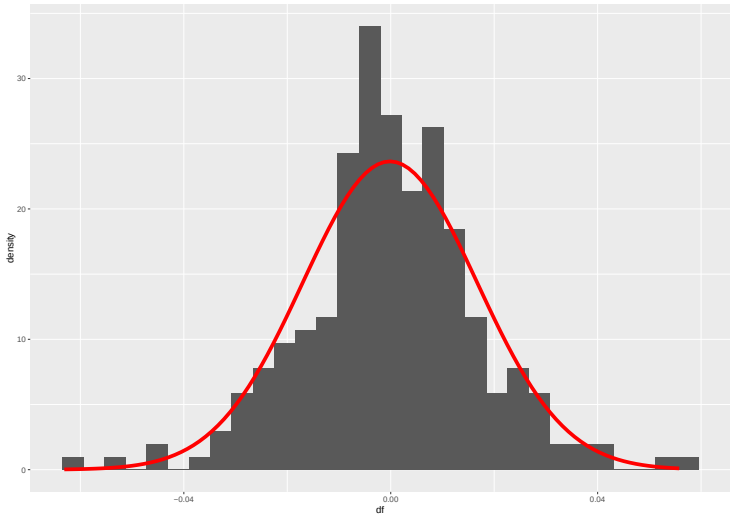

FIGURE 5 – Exercice 4