TD2 STATISTIQUES 2 / HPC - BIG DATA 2023

Exercice 1 :

Rappel :

La densité d'un vecteur aléatoire gaussien **Y** de dimension n, de matrice de variance-covariance inversible Γ et de vecteur espérance μ a pour expression :

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Gamma)}} \exp\left(-\frac{1}{2}\,{}^{t}(\mathbf{Y}-\mu)\Gamma^{-1}(\mathbf{Y}-\mu) \right)$$

On considère un modèle linéaire gaussien **hétéroscédastique** de dimension q, défini par :
$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, avec $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\,\boldsymbol{\Omega})$,
σ^2 constante inconnue et $\boldsymbol{\Omega}$ = diag(1,…,i,…,n), matrice diagonale.
La variance de la ième composante du vecteur **e** vaut dont $V[e_i] = i.\sigma^2$, pour i = 1,…,n.

1. Donner la loi suivie par le vecteur aléatoire **Y**.

   En déduire l'expression de la fonction de vraisemblance du vecteur **Y** notée $L(\boldsymbol{\beta},\sigma^2;\mathbf{Y})$ en fonction de $\|\mathbf{M}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\|^2$, **M** étant une matrice à définir (indépendante de σ^2).

2. En maximisant la log-vraisemblance $\ln(L(\boldsymbol{\beta},\sigma^2;\mathbf{Y}))$, montrer que les estimateurs du maximum de vraisemblance $\hat{\beta}_{MV}$ et $\hat{\sigma}^2_{MV}$ des paramètres $\boldsymbol{\beta}$ et σ^2 du modèle sont solutions du système :

   $$\|\mathbf{M}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\|^2 = n\sigma^2$$
   $${}^{t}\mathbf{X}\mathbf{M}^2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

   On considérera que la matrice ${}^{t}\mathbf{X}\mathbf{M}^2\mathbf{X}$ est définie positive.

3. En déduire les expressions de $\hat{\beta}_{MV}$ et $\hat{\sigma}^2_{MV}$. L'estimateur $\hat{\beta}_{MV}$ est-il biaisé ?

Exercice 2 :

On considère le modèle suivant :

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \qquad \text{pour } i = 1,\ldots,n$$

Les e_i sont supposés indépendantes, identiquement distribuées suivant une loi normale centrée de variance σ^2 constante et inconnue.

On suppose également que :

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \sum_{i=1}^{n} x_{ij}^{\,2} = n \qquad \text{pour } j = 1,2,3$$

$$\sum_{i=1}^{n} x_{i1}x_{i2} = 0 \qquad \sum_{i=1}^{n} x_{i1}x_{i3} = 0 \qquad \sum_{i=1}^{n} x_{i2}x_{i3} = n\theta \qquad \text{avec } -1 < \theta < 1$$

On note $S_o = \sum_{i=1}^{n} y_i$ et $S_j = \sum_{i=1}^{n} y_i x_{ij}$ pour $j = 1,2,3$

1. Matriciellement le modèle s'écrit : $\mathbf{Y} = \mathbf{T\beta} + \mathbf{e}$.
   Expliciter la matrice T et le vecteur $\boldsymbol{\beta}$. Quelle est la dimension q du modèle?

2. Ecrire les estimateurs des moindres carrés $\hat{\beta}_o$, $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\beta}_3$ explicitement en fonction de S_o, S_1, S_2, S_3, n et θ. Ces estimateurs sont-ils indépendants ?

3. Montrer que : $\|Y - T\hat{\beta}\|^2 = {}^tYY - {}^tYT\,({}^tTT)^{-1}\,{}^tTY$
   Montrer alors que : $\left\|Y - T\hat{\beta}\right\|^2 = \sum_{i=1}^{n} y_i^{\,2} - \frac{1}{n}\left(S_o^{\,2} + S_1^{\,2} + \frac{S_2^2 - 2\theta S_2 S_3 + S_3^{\,2}}{1-\theta^2} \right)$
   En déduire l'expression d'un estimateur non biaisé de σ^2.

4. Ecrire un test explicite (c'est-à-dire exprimé autant que possible à partir de quantités définies précédemment) de l'hypothèse nulle H_o : $\beta_1 = 0$ contre l'alternative H_1 : $\beta_1 \neq 0$ (donner : la statistique du test, sa loi sous H_o et la règle de décision au niveau α avec puis sans exploitation de la p_value).

Exercice 3 :

Soit Y une variable aléatoire expliquée avec la variable X par le modèle multiplicatif suivant :

$Y_i = e_i^* . \exp(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)$, avec i variant de 1 à 5. (**modèle 1**)

e_i^* est un terme d'erreur aléatoire strictement positif ; β_0, β_1 et β_2 les paramètres inconnus du modèle. On dispose d'une archive de 5 mesures Y_i du prédictand Y ainsi que des mesures correspondantes de la variable X : $X_1 = -2$, $X_2 = -1$, $X_3 = 0$, $X_4 = 1$ et $X_5 = 2$.

1. Proposer une fonction f de Y_i permettant de se ramener à un modèle linéaire gaussien classique (**modèle 2**). On suppose que les variables aléatoires réelles e_i , images des e_i^* par f, sont indépendantes et identiquement distribuées suivant la loi Normale $\mathcal{N}(0,\sigma^2)$, σ^2 étant un paramètre inconnu.

2. Matriciellement le **modèle 2** s'écrit : **f(Y)** = **Tβ** + **e**, où **f(Y)** (resp. **e**) est le vecteur de IR^5 de composantes $f(Y_i)$ (resp. e_i), et où **β** est le vecteur de IR^3 des paramètres du modèle. Expliciter la matrice **T** et le vecteur des paramètres **β**.

3. Expliciter la matrice **^{t}TT** puis calculer son inverse et l'exprimer sous la forme :

   $$({}^tTT)^{-1} = \frac{1}{70}\begin{pmatrix} a & 0 & d \\ 0 & b & 0 \\ d & 0 & c \end{pmatrix}$$ a, b, c et d étant 4 entiers relatifs à déterminer.

   Soit $\hat{\boldsymbol{\beta}}$ l'estimateur des moindres carrés de **β**. Exprimer les composantes $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$ du vecteur $\hat{\boldsymbol{\beta}}$ en fonction des données de l'archive. Expliciter la loi de $\hat{\beta}$ en fonction de β, σ^2 et T. Les différents estimateurs sont-ils indépendants ?

4. Donner l'expression de l'estimateur $\hat{\sigma}^2$ de σ^2 en fonction des données de l'archive.

5. On se demande si le terme quadratique du modèle a vraiment un intérêt. On souhaite alors tester au niveau α l'hypothèse H_o : « le coefficient β_2 est nul » contre l'hypothèse H_1 : « le coefficient β_2 est non nul ». Décrire le test envisagé : statistique du test, loi suivie par cette statistique sous H_0, règle de décision en exploitant la p-value du test.

6. Après estimation du **modèle 2**, celui-ci permettrait obtenir des estimations du prédictand f(Y) notées $f(Y_i)^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$. Comment déduiriez-vous simplement les estimations Y_i^* de la variable d'intérêt Y à partir des estimations $f(Y_i)^*$?

7. En exploitant les informations fournies ci-dessous, expliquer quel problème introduirait la démarche proposée à la question 6. Quel serait alors d'après vous la meilleure façon d'obtenir les estimations Y_i^* de la variable Y à partir du **modèle 2** ?

   Une variable aléatoire Z > 0 suit une loi log-normale de paramètres μ et σ^2 si la variable ln(Z) suit une loi normale de mêmes paramètres et on a les espérances suivantes : E[ln(Z)] =μ et E[Z]=exp(μ+ σ^2/2).

Exercice 4 :

On considère le modèle linéaire gaussien $\mathbf{Y = X\beta + e}$, avec $\mathbf{e \sim N_n(0, \sigma^2 I_n)}$, β étant un vecteur de dimension q et **X** une matrice de dimension (n,q), supposée de plein rang avec n>q.

On notera par la suite $\hat{\beta}$ et $\hat{\sigma}^2$ les estimateurs des moindres carrés respectivement de β et σ^2.

On définit le vecteur des résidus estimés $\boldsymbol{\varepsilon}$ par $\mathbf{Y = Y^* + \varepsilon}$, avec $\mathbf{Y^* = X}\hat{\beta}$ = vecteur des valeurs ajustées de Y par le modèle.

1. Rappeler sans démonstration les expressions des estimateurs des moindres carrés $\hat{\beta}$ et $\hat{\sigma}^2$.

2. Quelle est l'interprétation géométrique de $\mathbf{Y^*}$? Faire un schéma représentant **Y**, $\mathbf{Y^*}$ et $\boldsymbol{\varepsilon}$.

On notera par la suite Π_Q et $\Pi_{Q\perp}$ les matrices de projection orthogonale respectivement sur les sous-espaces vectoriels Q et $Q^{\perp}$.

3. Calculer les espérances et matrices de variance-covariance des vecteurs $\hat{\beta}$, $\mathbf{Y^*}$ et $\boldsymbol{\varepsilon}$.

4. Le modèle est de dimension 3 et on dispose de 13 observations du prédictand . On donne :

$$ {}^tXX = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 10 & 2 \\ 0 & 2 & 2 \end{pmatrix} \qquad {}^tXY = \begin{pmatrix} 10 \\ -5 \\ 3 \end{pmatrix} \qquad {}^tYY = 152.5 $$

Calculer $\hat{\beta}$, puis $\hat{\sigma}^2$.

5. En déduire une estimation de la matrice de variance-covariance de $\hat{\beta}$.
Les estimateurs des composantes de β sont-ils indépendants ?
Calculer le MSE des valeurs ajustées par ce modèle.

6. Le modèle exploite 2 prédicteurs X_1 et X_2 de la façon :
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}.X_{i2} + e_i$, avec i= 1,…,13.

Quelle sera alors l'impact, sur l'estimation du prédictand, d'une variation ΔX_1=+3 unités du prédicteur X_1 , le prédicteur X_2 restant constant et égal à 2 ?

Exercice 5 :

On dispose d'une série chronologique de n moyennes annuelles de pression atmosphérique, relative à une station météorologique. Les archives mentionnent de manière peu explicite un changement d'emplacement de la station de mesure à la fin de l'année k. On cherche à vérifier si ce changement a entraîné un biais significatif des mesures, dû à la modification éventuelle d'altitude du baromètre. Pour cela, on modélise la série temporelle de la manière suivante :

Soit Y_i la moyenne des pressions pour l'année i, avec i = 1,…,n.
On écrit le modèle sous la forme :

$Y_i = m + e_i$ pour i = 1,…,k
$Y_i = m + \delta + e_i$ pour i = k+1,…,n

Les paramètres m et δ sont inconnus. δ représente le biais éventuellement introduit dans les mesures suite au changement de site. Les résidus e_i sont supposés indépendants et identiquement distribués chacun selon la loi normale $N(0,\sigma^2)$, σ^2 étant un paramètre inconnu.

Matriciellement le modèle s'écrit : $\mathbf{Y = T\beta + e}$

On introduira les quantités suivantes : $\overline{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$ et $\overline{Y}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} Y_i$

1. Etablir les expressions, en fonction de $\overline{Y}_k$ et $\overline{Y}_{n-k}$, des estimateurs des moindres carrés $\hat{m}$ et $\hat{\delta}$ des paramètres m et δ du modèle.

2. Ces estimateurs sont-ils indépendants ? Donner leurs variances.

3. Etablir l'expression de l'estimateur $\hat{\sigma}^2$ de σ^2 en fonction de $\overline{Y}_k$ et $\overline{Y}_{n-k}$.

4. Proposer en le détaillant un test permettant de conclure sur la significativité d'un éventuel biais (Hypothèses H_0 et H_1, statistique du test, loi suivie sous H_0, règle de décision exploitant la p-value).
   Ce test a été réalisé sur les données et a mené à une p-value de 0.06. Conclure quant à la significativité du biais et commenter le résultat sachant que n=20 et k=15.