TD2 STATISTIQUES 2 / HPC - BIG DATA 2023

Modèle linéaire gaussien

Exercice 1 :

1)

L'homoscédasticité se caractérise par une variance conditionnelle constante de l'erreur.
Dans cet exercice, on modélise une situation hétéroscédastique.

$$\det(\Gamma) = \det(\sigma^2 \Omega) = n!\,\sigma^{2n}$$

$$L(\beta,\sigma^2;Y) = \frac{1}{\sqrt{n!}(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}{}^t(Y - X\beta)\Omega^{-1}(Y - X\beta)\right) = \frac{1}{\sqrt{n!}(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\Omega^{-\frac{1}{2}}(Y - X\beta)\right\|^2\right)$$

On a donc $M=\Omega^{-1/2}$

2)

Maximisons la Log-vraisemblance ln(L) , il vient :

$$\frac{\partial \ln L}{\partial \beta} = \frac{-1}{2\sigma^2}\frac{\partial}{\partial \beta}\left\|M(Y - X\beta)\right\|^2 = \frac{-1}{2\sigma^2}(-2{}^tXM^2(Y - X\beta)) \qquad \text{et} \qquad \frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\left\|M(Y - X\beta)\right\|^2$$

On obtient donc :

$$\left\|M(Y - X\hat{\beta}_{MV})\right\|^2 = n\hat{\sigma}^2_{MV} \qquad \text{et} \qquad {}^tXM^2(Y - X\hat{\beta}_{MV}) = 0$$

3)

Avec ${}^tXM^2X={}^tX\Omega^{-1}X$ inversible, on a alors :

$$\hat{\beta}_{MV} = ({}^tX\Omega^{-1}X)^{-1}\,{}^tX\Omega^{-1}Y \qquad \text{et} \qquad \hat{\sigma}^2_{MV} = \frac{1}{n}\left\|\Omega^{-\frac{1}{2}}(Y - X\hat{\beta}_{MV})\right\|^2$$

$$BIAIS[\hat{\beta}_{MV}] = E[\hat{\beta}_{MV}] - \beta = ({}^tX\Omega^{-1}X)^{-1}\,{}^tX\Omega^{-1}E[Y] - \beta = ({}^tX\Omega^{-1}X)^{-1}\,{}^tX\Omega^{-1}X\beta - \beta = 0$$

Exercice 2 :

1) Y=Tβ + e

Modèle de dimension q = 4 = dim(β), avec :

$$T=\begin{pmatrix}1 & x_{11} & x_{12} & x_{13}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2} & x_{n3}\end{pmatrix} \quad \text{et} \quad \beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}$$

2)

$$^tTT=\begin{pmatrix}n & 0 & 0 & 0\\ 0 & n & 0 & 0\\ 0 & 0 & n & n\theta\\ 0 & 0 & n\theta & n\end{pmatrix}=\begin{pmatrix}nI_2 & 0\\ 0 & A_{n,\theta}\end{pmatrix} \quad \text{et donc} \quad (^tTT)^{-1}=\begin{pmatrix}\frac{1}{n}I_2 & 0\\ 0 & A_{n,\theta}^{-1}\end{pmatrix}$$

avec $A_{n,\theta}^{-1}=\frac{1}{n(1-\theta^2)}\begin{pmatrix}1 & -\theta\\ -\theta & 1\end{pmatrix}$ et |θ|<1

On obtient alors pour les 4 estimateurs : $\hat{\beta}=(^tTT)^{-1}\,{}^tTY=\begin{pmatrix}\frac{1}{n}S_0\\ \frac{1}{n}S_1\\ \frac{1}{n(1-\theta^2)}(S_2-\theta S_3)\\ \frac{1}{n(1-\theta^2)}(S_3-\theta S_2)\end{pmatrix}$

On sait que la matrice de variance-covariance est : $V[\hat{\beta}]=\sigma^2(^tTT)^{-1}=\sigma^2\begin{pmatrix}\frac{1}{n}I_2 & 0\\ 0 & A_{n,\theta}^{-1}\end{pmatrix}$

Si θ=0, les covariances sont nulles et les 4 estimateurs gaussiens indépendants, mais si $\theta\neq 0$

$COV[\hat{\beta}_2,\hat{\beta}_3]=\sigma^2(^tTT)^{-1}_{34}=\frac{-\theta\sigma^2}{n(1-\theta^2)}\neq 0$ et ces 2 estimateurs ne sont alors pas indépendants.

3)

Avec les notations du cours :

$$\left\|Y-T\hat{\beta}\right\|^2=\left\|Y-\Pi_Q Y\right\|^2=\left\|\Pi_{Q^\perp}Y\right\|^2=\|\varepsilon\|^2$$

On obtient donc en développant :

$$\left\|Y-T\hat{\beta}\right\|^2={}^t(Y-\Pi_Q Y)(Y-\Pi_Q Y)={}^tYY-{}^tY\Pi_Q Y+{}^t(\Pi_Q Y)(\Pi_{Q^\perp}Y)={}^tYY-{}^tY\Pi_Q Y={}^tYY-{}^tYT(^tTT)^{-1}\,{}^tTY$$

Calculons $\quad {}^tYT({}^tTT)^{-1}\,{}^tTY = {}^tYT\hat{\beta}$

$$ {}^tYT\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i \left(S_0 + x_{i1}S_1 + \frac{x_{i2}(S_2 - \theta S_3)}{1-\theta^2} + \frac{x_{i3}(S_3 - \theta S_2)}{1-\theta^2} \right) = \frac{1}{n}\left(S_0^2 + S_1^2 + \frac{S_2^2 - 2\theta S_2 S_3 + S_3^2}{1-\theta^2} \right) $$

La relation demandée s'obtient alors aisément, on en déduit un estimateur non biaisé de σ^2 :

$$ \hat{\sigma}^2 = \frac{\left\| Y - T\hat{\beta} \right\|^2}{n-q} = \frac{1}{n-4}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(S_0^2 + S_1^2 + \frac{S_2^2 - 2\theta S_2 S_3 + S_3^2}{1-\theta^2} \right) \right] $$

4)

On veut tester, au niveau α, l'hypothèse nulle H_0 : « $\beta_1 = 0$ » contre l'hypothèse alternative H_1 : « $\beta_1 \neq 0$ » (→ test bilatéral, on ne sait rien sur le signe de β_1).
On a le résultat théorique suivant concernant l'estimateur de β_1 :

$\hat{\beta}_1 \qquad N(\beta_1, \sigma^2({}^tTT)^{-1}_{22}) \quad \sim \qquad N(\beta_1, \frac{\sigma^2}{n})$, soit la loi

La variable $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{n}}}$ suit alors une loi normale centrée réduite, et la variable $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{n}}}$ suivra une loi de Student à n-4 degrés de liberté, par passage de la variance théorique à la variance estimée.

Sous H_0 , la statistique $S = \dfrac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{n}}}$ suit donc la loi de Student à n – 4 degrés de liberté.

La règle de décision sera, en exploitant la p-value du test :

Si p_value < α : rejet de H_0 au niveau α.

Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α.

Sans la p-value, il faut comparer la valeur S_{obs} de la statistique de test prise sur vos données, (calculée en exploitant les formules établies aux questions 2 et 3) aux quantiles de la loi de Student à n-4 ddl qui définissent les zones de rejet du test, quantiles d'ordres $\alpha/2$ et $1-\alpha/2$ pour un test bilatéral de niveau α. La loi de Student étant symétrique, on a $q_{1-\alpha/2} = -\,q_{\alpha/2} > 0$

La règle de décision devient alors :

Si $|\,S_{obs}\,| > q_{1-\alpha/2}$: rejet de H_0 au niveau α.

Si $|\,S_{obs}\,| < q_{1-\alpha/2}$: les données ne permettent pas le rejet de H_0 au niveau α.

Exercice 3 :

1-

L'application de la fonction logarithme népérien permet de se ramener au cadre théorique du modèle linéaire gaussien. On obtient, avec i = 1,…,5 et $e_i = \ln(e_i^*)$, le modèle 2 suivant :

$$f(Y_i) = \ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

2-

f(Y) = **Tβ** + **e** , modèle de dimension q = 3 avec :

$$T = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

3-

$${}^tTT = \begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \quad \Rightarrow \quad ({}^tTT)^{-1} = \frac{1}{70}\begin{pmatrix} 34 & 0 & -10 \\ 0 & 7 & 0 \\ -10 & 0 & 5 \end{pmatrix}$$

$$\hat{\beta} = ({}^tTT)^{-1}\,{}^tTf(Y) = ({}^tTT)^{-1}\begin{pmatrix} \sum \ln(Y_i) \\ \sum X_i \ln(Y_i) \\ \sum X_i^2 \ln(Y_i) \end{pmatrix} = \frac{1}{70}\begin{pmatrix} \sum_i (34 - 10X_i^2)\ln(Y_i) \\ 7\sum_i X_i \ln(Y_i) \\ \sum_i (5X_i^2 - 10)\ln(Y_i) \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

$$\hat{\beta} \quad \sim \quad N_3(\beta, \sigma^2({}^tTT)^{-1}) \quad , \quad \text{et avec} \quad COV[\hat{\beta}_0, \hat{\beta}_2] = \frac{-\sigma^2}{7} \neq 0$$

➔ Ces 2 estimateurs gaussiens ne sont pas indépendants.

4-

$$\hat{\sigma}^2 = \frac{\left\| f(Y) - T\hat{\beta} \right\|^2}{5-3} = \frac{1}{2}\sum_{i=1}^{5}(\ln(Y_i) - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2)^2$$

5-

On veut tester, au niveau α, l'hypothèse nulle H_0 : « $\beta_2 = 0$ » contre l'hypothèse alternative H_1 : « $\beta_2 \neq 0$ » (test bilatéral). On a le résultat théorique suivant concernant l'estimateur de β_2 :

Sous H_0, la variable aléatoire

$$\frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_2}}} = \frac{\hat{\beta}_2}{\sqrt{({}^tTT)^{-1}_{33}.\hat{\sigma}^2}} = \frac{\hat{\beta}_2}{\sqrt{\frac{5}{70}.\frac{\left\|f(Y)-T\hat{\beta}\right\|^2}{2}}} = \frac{2\sqrt{7}\hat{\beta}_2}{\sqrt{\sum_i \left(\ln(Y_i) - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2\right)^2}}$$

suit une loi de Student à n-q=5-3=2 degrés de liberté.

La règle de décision est, en exploitant la p-value du test :

- Si p_value < α : rejet de H_0 au niveau α (le terme quadratique a un intérêt)
- Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α (terme quadratique sans influence).

6-

Il est tentant d'utiliser la fonction réciproque de f, et donc d'obtenir les estimations de la variable d'intérêt Y ainsi : $Y_i^* = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2)$.

7-

Dans le cadre théorique du modèle 2, **f(Y)** est un vecteur gaussien. D'après les informations fournies, si la variable ln(Y) suit une loi normale d'espérance $\beta_0 + \beta_1 X + \beta_2 X^2$ et de variance σ^2, la variable Y suit alors une loi log-normale de mêmes paramètres et on a :
$E[Y] = \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \sigma^2/2)$.

On voit donc que pour estimer Y, la démarche envisagée à la question 6 introduirait des erreurs potentiellement importantes et mènerait à sous-estimer l'espérance de Y (espérance conditionnelle aux prédicteurs). Par conséquent, il serait plus pertinent, pour déduire les estimations de Y à partir du modèle 2, de procéder ainsi :

$$Y_i^* = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \frac{\hat{\sigma}^2}{2}\right) = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \frac{1}{4}\sum_{i=1}^{5}\left(\ln(Y_i) - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2\right)^2\right)$$

Exercice 4 :

5)

$Y=X\beta + e$ avec $E[Y]=X\beta$ et $V[Y]=\sigma^2 I_n$

$$\hat{\beta} = ({}^tXX)^{-1}\,{}^tXY \qquad \hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-q}$$

6)

Y^*=vecteur de IR^n projeté orthogonal de Y sur Q, sous espace engendré par les vecteurs colonnes de X.
$\varepsilon=Y-Y^*$, vecteur des résidus estimés, et Y^*sont orthogonaux.

On a $Y^*=\Pi_Q Y$ avec $\Pi_Q=X({}^tXX)^{-1}\,{}^tX$ et $\varepsilon = Y - Y^* = (I_n - \Pi_Q)Y = \Pi_{Q^\perp} Y$

3)

$$E[\hat{\beta}] = ({}^tXX)^{-1}\,{}^tXE[Y] = ({}^tXX)^{-1}\,{}^tXX\beta = \beta \qquad V[\hat{\beta}] = ({}^tXX)^{-1}\,{}^tXV[Y]X({}^tXX)^{-1} = \sigma^2({}^tXX)^{-1}$$

$$E[Y^*] = XE[\hat{\beta}] = X\beta \qquad V[Y^*] = \Pi_Q V[Y]\,{}^t\Pi_Q = \sigma^2\Pi_Q^2 = \sigma^2\Pi_Q$$

$$E[\varepsilon] = E[Y] - E[Y^*] = X\beta - X\beta = 0 \qquad V[\varepsilon] = \Pi_{Q^\perp} V[Y]\,{}^t\Pi_{Q^\perp} = \sigma^2\Pi_{Q^\perp}^2 = \sigma^2\Pi_{Q^\perp}$$

4)

On obtient : $({}^tXX)^{-1} = \frac{1}{8}\begin{pmatrix} 8 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 5 \end{pmatrix}$ $\qquad \hat{\beta} = ({}^tXX)^{-1}\,{}^tXY = \begin{pmatrix} 10 \\ -1 \\ \frac{5}{2} \end{pmatrix}$

$$\hat{\sigma}^2 = \frac{\|\varepsilon\|^2}{n-q} = \frac{\|Y\|^2 - \|Y^*\|^2}{10}$$ (Pythagore)

$$\|Y^*\|^2 = {}^t(X\hat{\beta})X\hat{\beta} = {}^t(Y-\varepsilon)X\hat{\beta} = {}^tYX\hat{\beta} = 112.5$$

$$\hat{\sigma}^2 = \frac{152.5 - 112.5}{10} = 4$$

5)

Estimation de $V[\hat{\beta}]$: $\hat{V}[\hat{\beta}] = \hat{\sigma}^2({}^tXX)^{-1} = \frac{1}{2}\begin{pmatrix} 8 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 5 \end{pmatrix}$

Avec $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$, les estimateurs ne sont pas tous indépendants car $\hat{COV}[\hat{\beta}_1, \hat{\beta}_2] = -\frac{1}{2}$

$$MSE = \frac{1}{13}\sum_{i=1}^{13}\varepsilon_i^2 = \frac{\|\varepsilon\|^2}{13} = \frac{40}{13}$$

6)

Les valeurs ajustées par le modèle (estimations de Y calculées sur l'archive d'apprentissage) sont obtenues ainsi :
$Y_i^* = 10 - X_{i1} + 5/2\, X_{i1}.X_{i2}$

On a donc, si $X_2 = 2$:
$\Delta Y^* = -\Delta X_1 + 5\Delta X_1 = +12$ → Avec ce modèle, une augmentation de X_1 de 3 unités, X_2 restant égal à 2, entraînera une augmentation de 12 unités pour l'estimation de Y.

Exercice 5 :

1)

$Y = T\beta + e$, modèle de dimension q = 2 avec : (les 0 concernant les k premières lignes de T)

$$T = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \qquad \beta = \begin{pmatrix} m \\ \delta \end{pmatrix}$$

On obtient après calcul :

$$({}^tTT)^{-1} = \frac{1}{k}\begin{pmatrix} 1 & -1 \\ -1 & \dfrac{n}{n-k} \end{pmatrix} \qquad \hat{\beta} = ({}^tTT)^{-1}\,{}^tTY = \begin{pmatrix} \hat{m} \\ \hat{\delta} \end{pmatrix} = \begin{pmatrix} \overline{Y}_k \\ \overline{Y}_{n-k} - \overline{Y}_k \end{pmatrix}$$

2)

$$\hat{\beta} \sim N_2(\beta, \sigma^2 ({}^tTT)^{-1})$$

$$\hat{m} \sim N(m, \frac{\sigma^2}{k})$$

$$\hat{\delta} \sim N(\delta, \frac{n\sigma^2}{k(n-k)})$$

Ces 2 estimateurs ne sont pas indépendants, leur covariance étant non nulle (=-σ^2/k).

3)

$$\hat{\sigma}^2 = \frac{\left\| Y - T\hat{\beta} \right\|^2}{n-2} = \frac{1}{n-2}\left[\sum_{i=1}^{k} (Y_i - \hat{m})^2 + \sum_{i=k+1}^{n} (Y_i - \hat{m} - \hat{\delta})^2 \right] = \frac{1}{n-2}\left[\sum_{i=1}^{k} (Y_i - \overline{Y}_k)^2 + \sum_{i=k+1}^{n} (Y_i - \overline{Y}_{n-k})^2 \right]$$

$$= \frac{1}{n-2}\left[\sum_{i=1}^{n} Y_i^2 - k\overline{Y}_k^2 - (n-k)\overline{Y}_{n-k}^2 \right]$$

4)

On veut tester, au niveau α, l'hypothèse nulle H_0 : « $\delta = 0$ » contre l'hypothèse alternative H_1 : « $\delta \neq 0$ » (test bilatéral). On a le résultat théorique suivant concernant l'estimateur de δ :

Sous H_0, $\dfrac{\hat{\delta}}{\sqrt{\dfrac{n\hat{\sigma}^2}{k(n-k)}}}$ suit une loi de Student à n-2 degrés de liberté,

La règle de décision est, en exploitant la p-value du test :

Si p_value < α : rejet de H_0 au niveau α (existence d'un biais significatif)

Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α

(pas de biais significatif).

Ici p-value=0.06, en travaillant au niveau de risque 5%, le biais est alors jugé non significatif.

On peut cependant remarquer que la p-value est proche du seuil 0.05 et qu'on dispose d'un petit échantillon de 20 valeurs. De plus, on exploite uniquement 5 valeurs relatives à la station déplacée, avec un échantillon plus important (disponible dans quelques années) le biais pourrait devenir significatif.