TD4 STATISTIQUES 2 / HPC - BIG DATA 2023

Exercice 1 :

1)

Avec la contrainte '$\alpha_1 + \alpha_2 = 0$' imposée on obtient un modèle de dimension 4, la référence prise alors étant la demi-somme des impacts moyens des 2 modalités :

$$y_{ij} = (\mu + \frac{\alpha_1 + \alpha_2}{2}) + \beta_i x_{ij} + (\alpha_i - \frac{\alpha_1 + \alpha_2}{2}) + e_{ij} \quad \text{, soit :} \quad y_{ij} = \mu' + \beta_i x_{ij} + \alpha_i' + e_{ij}$$

Avec $\mu' = \mu + (\alpha_1 + \alpha_2)/2$ et $\alpha_i' = \alpha_i - (\alpha_1 + \alpha_2)/2$, on a alors :

$\alpha_1' = (\alpha_1 - \alpha_2)/2$ et $\alpha_2' = (\alpha_2 - \alpha_1)/2 = -\alpha_1'$

$Y = T\beta + e$ avec :

$$\beta = \begin{pmatrix} \mu' \\ \alpha_1' \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad T = \begin{pmatrix} 1 & 1 & -2 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & -1 & 0 & -2 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 2 \end{pmatrix} \qquad {}^tTT = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{pmatrix}$$

$$({}^tTT)^{-1} = \frac{1}{10} I_4$$

2)

$$\hat{\beta} = ({}^tTT)^{-1}\,{}^tTY = \frac{1}{10}{}^tTY = \frac{1}{10}\begin{pmatrix} \sum_{ij} y_{ij} \\ \sum_j (y_{1j} - y_{2j}) \\ \sum_j x_{1j} y_{1j} \\ \sum_j x_{2j} y_{2j} \end{pmatrix}$$

La matrice de variance-covariance de ce vecteur $\frac{\sigma^2}{10} I_4$ étant diagonale, les 4 estimateurs sont indépendants.

On a vu à la question précédente que le paramètre estimé relatif au facteur F est, avec la contrainte imposée, le demi-effet différentiel : $\alpha_1' = (\alpha_1 - \alpha_2)/2$

3)

$$\hat{\beta} \quad \sim \quad N_4(\beta, \frac{\sigma^2}{10} I_4)$$

Les estimateurs étant indépendants, on a : $V[\hat{\beta}_1 - \hat{\beta}_2] = V[\hat{\beta}_1] + V[\hat{\beta}_2] = \frac{\sigma^2}{5}$

Attention, si non indépendance on a : $V[\hat{\beta}_1 - \hat{\beta}_2] = V[\hat{\beta}_1] + V[\hat{\beta}_2] - 2COV[\hat{\beta}_1, \hat{\beta}_2]$

4)

On veut tester, au niveau α, l'hypothèse nulle H_0 : « $\beta_1 - \beta_2 = 0$ » contre l'hypothèse alternative H' : « $\beta_1 - \beta_2 \neq 0$ » (test bilatéral).

On a le résultat théorique suivant concernant l'estimateur de $\beta_1 - \beta_2$:

$\hat{\beta}_1 - \hat{\beta}_2 \quad \sim \quad N(\beta_1 - \beta_2, \frac{\sigma^2}{5})$ (estimateurs gaussiens non biaisés)

La variable $\dfrac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{\sqrt{\dfrac{\sigma^2}{5}}}$ suit alors une loi normale centrée réduite.

Et sous H_0, la variable $\dfrac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\dfrac{\hat{\sigma}^2}{5}}}$ suivra donc une loi de Student à 10 - 4 = 6 ddl.

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

Si p_value < α : rejet de H_0 au niveau α (on considère que $\beta_1 \neq \beta_2$)

Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α ($\beta_1 = \beta_2$).

Exercice 2 :

1)

Le modèle à rupture s'écrit (pour i=1 à 20, FF_i = i m/s) :

$$
\begin{aligned}
&FF_i \le 5 : P_i = \beta_0 + e_i \\
&5 \le FF_i \le 15 : P_i = \beta_1 + \beta_2 FF_i + e_i \\
&FF_i \ge 15 : P_i = \beta_3 + e_i
\end{aligned}
$$

Les contraintes de continuité en 5 et 15 m/s font que sa dimension est égale à 2 car :

$$
\begin{aligned}
&\beta_0 = \beta_1 + 5\beta_2 \\
&\beta_3 = \beta_1 + 15\beta_2
\end{aligned}
$$

Le modèle est donc défini par :

$$
\begin{aligned}
&FF_i \le 5 : P_i = \beta_1 + 5\beta_2 + e_i \\
&5 \le FF_i \le 15 : P_i = \beta_1 + \beta_2 FF_i + e_i \\
&FF_i \ge 15 : P_i = \beta_1 + 15\beta_2 + e_i
\end{aligned}
$$

2)

P = Mβ + e , avec :

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$ {}^tM = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 5 & 5 & 5 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 15 & 15 & 15 & 15 & 15 \end{pmatrix}$$

$$ {}^tMM = \begin{pmatrix} 20 & 205 \\ 205 & 2435 \end{pmatrix} \qquad ({}^tMM)^{-1} = \frac{1}{1335}\begin{pmatrix} 487 & -41 \\ -41 & 4 \end{pmatrix}$$

On obtient alors pour les 2 estimateurs : $\hat{\beta} = ({}^tMM)^{-1}\,{}^tMP = \dfrac{1}{1335}\begin{pmatrix} 487\sum_{i=1}^{20} P_i - 41(5\sum_{i=1}^{5} P_i + \sum_{i=6}^{15} iP_i + 15\sum_{i=16}^{20} P_i) \\ -41\sum_{i=1}^{20} P_i + 4(5\sum_{i=1}^{5} P_i + \sum_{i=6}^{15} iP_i + 15\sum_{i=16}^{20} P_i) \end{pmatrix}$

3)

On sait que la matrice de variance-covariance est : $V[\hat{\beta}] = \sigma^2 ({}^t MM)^{-1} = \dfrac{\sigma^2}{1335}\begin{pmatrix} 487 & -41 \\ -41 & 4 \end{pmatrix}$

$\hat{\beta}_1 \sim N(\beta_1, \dfrac{487\sigma^2}{1335})$

$\hat{\beta}_2 \sim N(\beta_2, \dfrac{4\sigma^2}{1335})$

$COV[\hat{\beta}_1, \hat{\beta}_2] = \sigma^2 ({}^t MM)^{-1}_{12} = \dfrac{-41\sigma^2}{1335} \neq 0$, ces 2 estimateurs ne sont donc pas indépendants.

4)

En considérant les matrices de projections orthogonales définies en cours, on obtient :

$$\left\| P - M\hat{\beta} \right\|^2 = \left\| P - \Pi_Q P \right\|^2 = {}^t(P - \Pi_Q P)(P - \Pi_Q P) = {}^t PP - {}^t P \Pi_Q P + {}^t(\Pi_Q P)(\Pi_{Q^\perp} P) = {}^t PP - {}^t PM\hat{\beta}$$

$$\hat{\sigma}^2 = \frac{\left\| P - M\hat{\beta} \right\|^2}{n-q} = \frac{{}^t PP - {}^t PM\hat{\beta}}{18} = \frac{1}{18}\left(\sum_{i=1}^{20} P_i^2 - {}^t PM\hat{\beta}\right)$$

5)

L'éolienne 2 aura de meilleures performances que l'éolienne 1 si sa réponse au vent est plus importante, donc si $\beta'_2 > \beta_2$, en notant β'_2 le second paramètre du modèle à rupture relatif à la seconde éolienne.

Le modèle 2 est défini par P'=M'β'+e'.

Les matrice M et M' sont égales puisque les tests des éoliennes ont été réalisés dans les mêmes conditions (mêmes valeurs du prédicteur FF). De plus, les variances des erreurs des 2 modèles sont supposées égales ainsi que leurs estimations.
Les matrices de variance-covariance des estimateurs des vecteurs β et β' sont donc identiques et égales à $\sigma^2({}^t MM)^{-1}$.

Les tests des éoliennes ayant été menés de manière indépendante, les estimateurs relatifs au modèle 1 sont indépendants des estimateurs relatifs au modèle 2, on a donc :

$$V[\hat{\beta}_2' - \hat{\beta}_2] = V[\hat{\beta}_2'] + V[\hat{\beta}_2] = 2V[\beta_2] = \frac{8\sigma^2}{1335}$$

On veut tester, au niveau α, l'hypothèse nulle H_0 : « $\beta'_2 - \beta_2 = 0$ » contre l'hypothèse alternative H' : « $\beta'_2 - \beta_2 \neq 0$ » (test bilatéral).

On a le résultat théorique suivant concernant l'estimateur de $\beta'_2 - \beta_2$:

$$\hat{\beta}'_2 - \hat{\beta}_2 \quad \sim \quad N(\beta'_2 - \beta_2, \frac{8\sigma^2}{1335})$$ (estimateurs gaussiens non biaisés)

La variable $\dfrac{(\hat{\beta}'_2 - \hat{\beta}_2) - (\beta'_2 - \beta_2)}{\sqrt{\dfrac{8\sigma^2}{1335}}}$ suit alors une loi normale centrée réduite.

Et sous H_0, la variable $\dfrac{\hat{\beta}'_2 - \hat{\beta}_2}{\sqrt{\dfrac{8\hat{\sigma}^2}{1335}}}$ suivra donc une loi de Student à 20 - 2 = 18 ddl.

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

Si p_value < α : rejet de H_0 au niveau α (on considère que $\beta'_2 \neq \beta_2$)

Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α ($\beta'_2 = \beta_2$).

La seconde éolienne aura donc de meilleures performances que la première si :

p_value < α et si la valeur prise par la statistique de test sur l'échantillon est positive

(et dans ce cas on a bien $\beta'_2 > \beta_2$) , sinon (statistique négative) on conclura que c'est l'éolienne 1 qui est la plus performante.

Remarque : un test unilatéral aurait été effectué si on avait la certitude que l'éolienne 1 ne peut pas être plus performante que l'éolienne 2 (= performances identiques ou éolienne 2 meilleure).

Exercice 3 :

1-

Dans le cadre d'une régression logistique, le logit de la probabilité de succès, conditionnelle aux prédicteurs, est modélisé linéairement. On obtient donc la modélisation suivante :

$$P(Y=1|X)=\frac{e^{{}^t\beta\,X}}{1+e^{{}^t\beta\,X}}$$

2-

Les mesures du prédictand étant indépendantes, la fonction de vraisemblance de l'échantillon, fonction des k paramètres inconnus β_j , est définie par (avec y_i = 0 ou 1) :

$$L(\beta)=\prod_{i=1}^{n}P(Y=y_i|X_i)=\prod_{i=1}^{n}\left(\frac{e^{{}^t\beta\,X_i}}{1+e^{{}^t\beta\,X_i}}\right)^{y_i}\left(1-\frac{e^{{}^t\beta\,X_i}}{1+e^{{}^t\beta\,X_i}}\right)^{1-y_i}=\prod_{i=1}^{n}\frac{e^{y_i\,{}^t\beta\,X_i}}{1+e^{{}^t\beta\,X_i}}=\frac{e^{\sum_{i=1}^{n}y_i\,{}^t\beta\,X_i}}{\prod_{i=1}^{n}(1+e^{{}^t\beta\,X_i})}$$

3-

La log-vraisemblance a pour expression $\ln(L(\beta))=\sum_{i=1}^{n}[y_i\,{}^t\beta\,X_i-\ln(1+e^{{}^t\beta\,X_i})]$

L'estimateur cherché de β vérifie alors :

$$\left(\frac{\partial\ln(L(\beta))}{\partial\beta}\right)_{\beta=\hat\beta}=0 \quad \text{or} \quad \frac{\partial\ln(L(\beta))}{\partial\beta}=\sum_{i=1}^{n}\left(y_iX_i-\frac{e^{{}^t\beta\,X_i}}{1+e^{{}^t\beta\,X_i}}X_i\right)$$

L'estimateur vérifie l'équation vectorielle suivante, équivalente au système à k équations de l'énoncé :

$$\sum_{i=1}^{n}y_iX_i=\sum_{i=1}^{n}\frac{e^{{}^t\hat\beta\,X_i}}{1+e^{{}^t\hat\beta\,X_i}}X_i \quad\Leftrightarrow\quad \sum_{i=1}^{n}y_ix_{ij}=\sum_{i=1}^{n}\frac{e^{{}^t\hat\beta\,X_i}}{1+e^{{}^t\hat\beta\,X_i}}x_{ij}$$, avec j=1,…,k.