TD4 STATISTIQUES 2 / HPC - BIG DATA 2023

Exercice 1 :

On veut modéliser l'influence d'un prédicteur qualitatif (facteur) F à deux modalités nommées i = 1 ou 2 et d'un prédicteur quantitatif X sur une variable réponse quantitative Y. On dispose des mesures suivantes :

(i,j)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)
X	-2	-1	0	1	2	-2	-1	0	1	2
F	1	1	1	1	1	2	2	2	2	2
Y	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}

On considère alors le modèle linéaire suivant :

$$y_{ij} = \mu + \alpha_i + \beta_i . x_{ij} + e_{ij} \qquad \text{pour } i=1 \text{ à } 2 \quad j=1,\dots,5$$

Où Y_{ij} est la variable réponse, α_i l'effet fixe relatif à la $i^{ème}$ modalité du facteur F et e_{ij} la variable erreur. Le coefficient de régression β_i prend deux valeurs β_1 ou β_2 en fonction de la modalité de la variable F.

On impose la contrainte d'identification des paramètres α_i suivante : « $\alpha_1 + \alpha_2 = 0$ »

Les variables aléatoires réelles e_{ij} sont supposées indépendantes et identiquement distribuées suivant la loi Normale $\boldsymbol{N}(0,\sigma^2)$, σ^2 étant un paramètre inconnu.

Matriciellement le modèle s'écrit :

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\beta} + \mathbf{e}$$

Où **Y** (resp. e) est le vecteur de IR^{10} de composantes y_{ij} (resp. e_{ij}) sur la base canonique, et où $\boldsymbol{\beta}$ est le vecteur de IR^q des paramètres du modèle dans la base canonique.

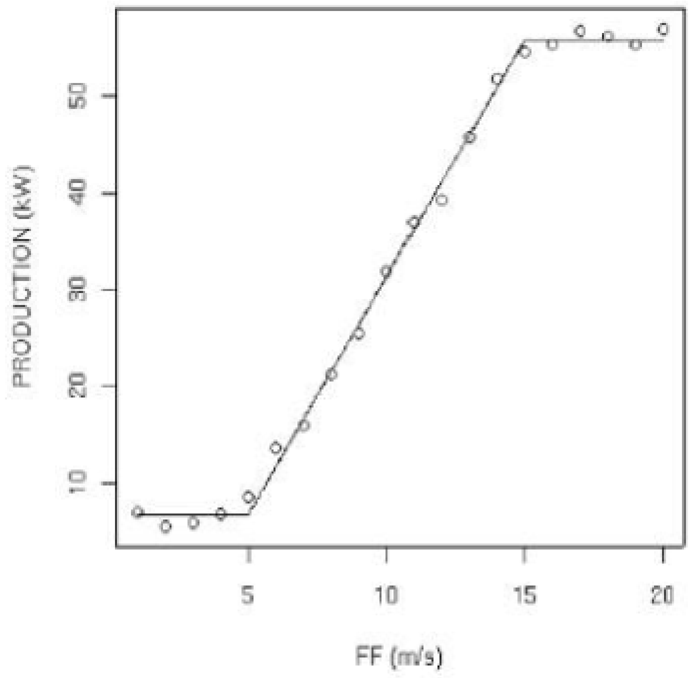
1. Expliciter, en considérant la contrainte imposée, la matrice **T** et le vecteur $\boldsymbol{\beta}$.
   Quelle est la dimension q du modèle? Expliciter la matrice ${}^t\mathbf{TT}$ et calculer son inverse.

2. Exprimer les estimations des composantes du vecteur $\boldsymbol{\beta}$ en fonction des y_{ij} et x_{ij}.
   Les estimateurs sont-ils indépendants ? Comment interpréter les paramètres estimés ?

3. Expliciter la loi de $\hat{\beta}$ en fonction de $\boldsymbol{\beta}$, σ^2 et **T**. En déduire la variance et l'espérance de la variable aléatoire $\hat{\beta}_1-\hat{\beta}_2$.

4. En déduire un test de niveau α de l'hypothèse $H_o : \beta_1-\beta_2=0$ contre $H_1 : \beta_1-\beta_2\neq 0$
   (Statistique du test, sa loi sous H_0 et la règle de décision en exploitant la p-value).

Exercice 2 :



La maquette d'un nouveau type d'éolienne est testée en soufflerie. 20 mesures sont réalisées, de 1 à 20 m/s par pas de 1 m/s. L'allure de la réponse (graphe ci-dessus) suggère un modèle à rupture.
La production P est modélisée en fonction du vent FF généré dans la soufflerie de la façon suivante : entre 1 et 5 m/s la réponse est supposée constante, elle augmente linéairement entre 5 et 15 m/s, avant de saturer au-delà de 15 m/s. Il y a continuité de la réponse P aux points correspondant à 5 et 15 m/s.

1)
Ecrire le modèle liant les Y_i aux FF_i en fonction des plages de valeurs de FF. Montrer que du fait de l'hypothèse de continuité de la réponse P, le modèle proposé est de dimension 2.

2)
Le modèle s'écrit matriciellement $P = M\beta + e$, où P (resp. e) est le vecteur de IR^{20} de composantes P_i (resp. e_i) sur la base canonique, et où β est le vecteur de IR^2 des paramètres du modèle. Les variables aléatoires réelles e_i sont supposées indépendantes, gaussiennes, centrées de variances constante et inconnue σ^2. Expliciter la *design matrix* M et le vecteur β.
Donner les expressions des estimateurs des moindres carrés des 2 paramètres du modèle en fonction des P_i.

3)
Donner les lois de ces 2 estimateurs. Sont-ils indépendants ?

4)
Montrer qu'un estimateur sans biais de la variance du terme d'erreur s'écrit : $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{20} P_i^2 - {}^tPM\hat{\beta}}{18}$

5)
Un deuxième type d'éolienne est testé, par la même procédure expérimentale, indépendamment du premier. La réponse de cette éolienne suit le même type de modèle. On suppose que les performances des 2 éoliennes sont proches entre 1 et 5 m/s, mais pour des valeurs de vent supérieures à 5 m/s la deuxième éolienne semble avoir de meilleures performances. On veut vérifier si cette différence constatée est significative au moyen d'un test statistique.

Sur quel paramètre du modèle doit porter ce test ?
Ce paramètre est estimé pour chacune des éoliennes. Pour simplifier on suppose que les variances d'erreurs sont les mêmes pour les 2 éoliennes, ainsi que leurs estimations.
Donner la loi suivie par la différence des 2 estimateurs. En déduire un test de niveau α permettant de tester si la deuxième éolienne est plus performante que la première.

Exercice 3 : régression logistique

On considère une variable de Bernoulli Y expliquée par un modèle de régression logistique sans terme constant et exploitant k prédicteurs. Le vecteur $\boldsymbol{\beta}$ de IR^k a pour composantes les k paramètres inconnus β_j du modèle logistique et le vecteur aléatoire **X** de IR^k, les k prédicteurs: ${}^t\mathbf{X} = (X_1 \dots X_k)$.

1) Exprimer P(Y=1| **X**), la probabilité de succès conditionnelle aux prédicteurs, en fonction des vecteurs $\boldsymbol{\beta}$ et **X**.

2) On dispose d'une archive de n mesures du prédictand Y, ainsi que des mesures correspondantes des k prédicteurs. On note y_i la ième mesure de la variable Y et $\mathbf{X_i}$ le vecteur de IR^k de composantes les ièmes mesures des k prédicteurs :
${}^t\mathbf{X}_i = (x_{i1} \dots x_{ik})$.
Les mesures y_i étant indépendantes, donner l'expression de la fonction de vraisemblance de l'échantillon $(y_1,\dots,y_n)$ en fonction des y_i , des vecteurs $\mathbf{X_i}$ et du vecteur $\boldsymbol{\beta}$.

3) Montrer que $\hat{\boldsymbol{\beta}}$, estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$, vérifie le système constitué des k équations suivantes :

$$\sum_{i=1}^{n} \frac{e^{{}^t\hat{\beta} X_i}}{1+e^{{}^t\hat{\beta} X_i}} x_{ij} = \sum_{i=1}^{n} y_i x_{ij} \qquad \text{, avec } j=1,\dots,k.$$