TD3 STATISTIQUES 2 / HPC - BIG DATA 2023

Exercice 1 :

On cherche à expliquer la production photovoltaïque (PROD en W) par deux prédicteurs que sont les paramètres rayonnement (RAY en W/m^2) et température de l'air (TEMP en °C). On dispose de 30 mesures de chacune des variables, les mesures de production photovoltaïque étant issues de trois capteurs de surface 1 m^2 provenant de constructeurs différents nommés a, b et c (voir tableau ci-dessous).

i	1	…	10	11	…	20	21	…	30
PROD	P_1				…				P_{30}
RAY	R_1				…				R_{30}
TEMP	T_1				…				T_{30}
CAPT	a	…	a	b	…	b	c	…	c

On considère le modèle suivant :

$$P_i = \beta_o + \beta_1 R_i + \beta_2 T_i + e_i \qquad \text{pour } i = 1,\ldots, 30 \qquad \text{(modèle 1)}$$

Les e_i sont supposées indépendantes, identiquement distribuées suivant une loi normale centrée de variance σ^2 constante et inconnue. Les variables RAY et TEMP sont **centrées**.

De plus, on donne : $\sum_{i=1}^{30} T_i^2 = 10^4$; $\sum_{i=1}^{30} R_i^2 = 10^6$; $\sum_{i=1}^{30} R_i T_i = 0$

1. Matriciellement le modèle s'écrit : $\mathbf{P} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.
   Expliciter la matrice $\mathbf{X}$ et le vecteur $\boldsymbol{\beta}$. Quelle est la dimension q du modèle ?
   Donner l'expression des estimateurs des moindres carrés $\hat{\beta}_o$, $\hat{\beta}_1$, $\hat{\beta}_2$ et expliciter la loi de chacun d'eux. Ces estimateurs sont-ils indépendants ? Donner l'expression de l'estimateur sans biais de σ^2.

2. Proposer formellement (poser les hypothèses H_o et H_1, la statistique de décision, sa loi sous H_o, la règle de décision) un test de niveau α permettant de tester la significativité de l'effet de la température de l'air sur la production photovoltaïque.

3. Au vu des résultats ci-dessous obtenus avec le logiciel R sur les données du problème, répondre aux questions suivantes :

   a. Donner les valeurs de $\hat{\beta}_1$ et $\hat{\beta}_2$. Peut-on conclure que la température de l'air a un effet significatif sur la production photovoltaïque ?
      Qu'en est-il du prédicteur rayonnement global ? Justifier les réponses.
   b. Calculer la donnée manquante notée « ? » dans les sorties R fournies.
   c. Donner une estimation de la variance du terme d'erreur dans ce modèle.
   d. Interpréter le nombre *Multiple R-Squared* fourni.

lm(formula = PROD ~ RAY + TEMP, data)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	***80.50***	***2.00***	***?***	***1.09e-12***
RAY	***0.043***	***0.01***	***4.30***	***0.00021***
TEMP	***-0.004***	***0.11***	***-0.036***	***0.972***

Residual standard error: 10.95 on 27 degrees of freedom
Multiple R-squared: 0.6865

4. On cherche désormais à répondre à la question suivante :
les trois capteurs photovoltaïques ont-ils des comportements différents ?

   On introduit dans notre modèle le paramètre α_{capt} pouvant prendre trois valeurs (α_a, α_b et α_c) en fonction de la modalité du facteur CAPT (voir tableau page 1). On considère donc le modèle d'analyse de covariance suivant :

   $P_i = \beta_o + \beta_1 R_i + \beta_2 T_i + (\alpha_{capt})_i + e_i$ pour i = 1,…, 30 (modèle 2)

   Les e_i sont supposées indépendantes, identiquement distribuées suivant une loi normale centrée de variance σ^2 constante et inconnue. Le modèle s'écrit : $\mathbf{P} = \mathbf{Z}\beta + \mathbf{e}$.

   Expliciter la matrice $\mathbf{Z}$ et le vecteur β du modèle 2 en imposant la contrainte d'identification notée « α_a=0 » . Quelle est la dimension q de ce modèle ? Avec cette contrainte, comment interpréter les coefficients qui seront estimés ?

   Les sorties R listées ci-dessous et obtenues sur les données du problème permettent-elles de répondre à la question posée ? Justifier.

lm(formula = PROD ~ RAY + TEMP + CAPT, data)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	***73.86***	***1.75***	***42.21***	***1.15e-13***
RAY	***0.043***	***0.008***	***5.375***	***1.41e-05***
TEMP	***-0.004***	***0.10***	***-0.040***	***0.9681***
$\boldsymbol{CAPT_b}$	***3.21***	***1.53***	***2.098***	***0.0462***
$\boldsymbol{CAPT_c}$	***0.11***	***1.53***	***0.072***	***0.9432***

Exercice 2 :

On cherche à prévoir pour le lendemain la valeur d'un indice noté PO de pollution à l'ozone en exploitant trois prédicteurs potentiels constitués de prévisions d'un modèle météorologique à l'échéance 24H des variables suivantes :

- T, la température T de l'air en °C.
- FF, la force du vent en m/s.
- DD, la direction du vent (**variable qualitative à 4 modalités : Nord, Ouest, Sud et Est**).

L'indice PO est défini par : **PO = [O_3] / 180**, avec **[O_3]** concentration en Ozone en $\mu g.m^{-3}$, 180 $\mu g.m^{-3}$ correspondant au seuil de concentration au-delà duquel la population doit être informée.

On dispose d'une archive de **n = 80** valeurs de chacune des variables, prédictand et prédicteurs :

i	**1**	**…**	**20**	**21**	**…**	**40**	**41**	**…**	**60**	**61**	**…**	**80**
PO	PO_1				…						…	PO_{80}
T	T_1				…						…	T_{80}
FF	FF_1				…						…	FF_{80}
DD	Nord	…	Nord	Ouest	…	Ouest	Sud	…	Sud	Est	…	Est

<u>Partie I</u> : Régression multiple

On souhaite élaborer puis tester le modèle de régression suivant, comportant un terme d'interaction proportionnel au produit des variables T et FF :

$$PO_i = \beta_0 + \beta_1 T_i + \beta_2 FF_i + \beta_3 FF_i * T_i + e_i \qquad \text{pour } i = 1,\ldots, n \qquad \textbf{(modèle 1)}$$

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

Les variables FF et T ont été centrées et normées : $\sum_{i=1}^{n} T_i = \sum_{i=1}^{n} FF_i = 0$ **et** $\sum_{i=1}^{n} T_i^2 = \sum_{i=1}^{n} FF_i^2 = n$ **. De plus, on donne :** $\sum_{i=1}^{n} T_i.FF_i^2 = -40$ **et** $\sum_{i=1}^{n} T_i^2.FF_i^2 = +40$

Les sommes $\sum_{i=1}^{n} T_i.FF_i$ **et** $\sum_{i=1}^{n} T_i^2.FF_i$ **, négligeables devant les autres sommes du problème, seront considérées comme nulles.**

1. Matriciellement le modèle 1 s'écrit : **PO** = **X**β + **e**.
   Entre-t-il dans le cadre théorique du modèle linéaire gaussien ? Pourquoi ?
   Quelles hypothèses sont alors faites ? Quelle est la dimension q du modèle 1 ?
   Donner les expressions de la *'design matrix'* **X** et du vecteur des paramètres β.

2. Donner l'expression de la matrice ${}^tXX$ puis exprimer son inverse sous la forme :

$$({}^tXX)^{-1} = \frac{1}{80}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & a & a \\ 0 & 0 & a & b \end{pmatrix},$$ a et b étant 2 entiers à déterminer.

3. Donner les expressions des estimateurs des moindres carrés des paramètres β_0, β_1, β_2 et β_3. Expliciter la loi du vecteur $\hat{\beta}$.
Les estimateurs des paramètres sont-ils indépendants ?

4. Proposer formellement (poser les hypothèses nulle et alternative, la statistique du test, sa loi sous H_0, la règle de décision en exploitant la p-value) un test de niveau α permettant de tester la significativité de l'effet de la force du vent sur l'indice PO.

5. Au vu des résultats ci-dessous obtenus avec le logiciel R sur les données du problème, répondre aux questions suivantes :

   a. Donner la valeur de $\hat{\beta}_3$.
      Le terme d'interaction du modèle présente-t-il un intérêt ?
   b. Quel est le pourcentage de variance expliquée par ce modèle ?
   c. Rappeler l'expression de l'estimateur sans biais de σ^2, la variance du terme d'erreur du modèle. Donner la valeur numérique de l'estimation de l'écart-type $\hat{\sigma}$ de l'erreur.
   d. Quelle est l'unité de $\hat{\beta}_1$? Quel est l'écart-type estimé de cet estimateur ?
      Comment décririez-vous l'effet moyen de la température sur l'indice PO ?
   e. Au final, quels prédicteurs retenir ? Justifier la réponse.

Call:
lm(formula = PO ~ T + FF + FF * T, data)

Coefficients:

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	***0.75***	***0.016***	***46.8***	***< 2e-16***
T	***0.15***	***0.016***	***9.37***	***1.19e-14***
FF	***-0.06***	***0.023***	***-2.61***	***0.000782***
T*FF	***-0.01***	***0.033***	***-0.30***	***0.547740***

Residual standard error: 0.15 on 76 degrees of freedom
Multiple R-squared: 0.60

Partie II : Analyse de covariance

Afin de tester l'intérêt du prédicteur DD, on considère le modèle d'analyse de covariance suivant :

$$PO_i = \beta_o + \beta_1 T_i + \beta_2 FF_i + \theta_i + e_i \qquad \text{pour } i = 1,\ldots, n \qquad \textbf{(modèle 2)}$$

θ_i pouvant prendre 4 valeurs : θ_N, θ_O, θ_S et θ_E suivant la modalité du facteur DD.
Les erreurs e_i sont supposées indépendantes, identiquement distribuées suivant une loi normale centrée de variance σ^2 constante et inconnue.

1. Matriciellement le modèle 2 s'écrit : $\mathbf{PO} = \mathbf{Z}\beta + \mathbf{e}$.
   En imposant la contrainte d'identification notée « $\theta_E = 0$ », expliciter $\mathbf{Z}$ et β.
   Quelle est la dimension q du modèle 2 ?
   Avec la contrainte imposée, comment interprétez-vous les paramètres estimés ?

2. A l'aide des sorties R listées ci-dessous et obtenues sur les données du problème, répondre aux questions suivantes :

   a. Interpréter les résultats relatifs au prédicteur DD.
   Est-il pertinent de conserver cette variable dans le modèle ? Justifier la réponse.
   b. Peut-on dire que la modalité Ouest du facteur DD est sans effet sur le prédictand ?
   c. Le modèle 2 est-il meilleur que le modèle 1 ? Justifier la réponse.
   d. Les modèles testés vous semblent-ils adaptés au prédictand étudié ?
   Justifier la réponse.

Call:
lm(formula = PO ~ T + FF + DD, data)

Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
-0.28335	*-0.07115*	*-0.01171*	*0.08507*	*0.27618*

Coefficients:

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*0.71*	*0.026*	*27.31*	*< 2e-16*
T	*0.14*	*0.013*	*10.77*	*< 2e-16*
FF	*-0.05*	*0.013*	*-3.84*	*0.00042*
DDN	*0.18*	*0.037*	*4.86*	*3.99e-06*
DDO	*0.02*	*0.038*	*0.53*	*0.59791*
DDS	*-0.02*	*0.037*	*-0.54*	*0.45814*

Residual standard error: 0.11 on 74 degrees of freedom
Multiple R-squared: 0.74