TD3 STATISTIQUES 2 / HPC - BIG DATA 2023

Modèle linéaire gaussien - Prédicteurs qualitatifs

Exercice 1 :

1)

P = Xβ + e , modèle de dimension q = 3 avec :

$$X = \begin{pmatrix} 1 & R_1 & T_1 \\ \vdots & \vdots & \vdots \\ 1 & R_{30} & T_{30} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$^tXX = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 10^6 & 0 \\ 0 & 0 & 10^4 \end{pmatrix} \qquad \hat{\beta} = (^tXX)^{-1}{}^tXP = \begin{pmatrix} \frac{1}{30}\sum_{i=1}^{30} P_i \\ 10^{-6}\sum_{i=1}^{30} R_i P_i \\ 10^{-4}\sum_{i=1}^{30} T_i P_i \end{pmatrix}$$

$$\hat{\beta} \sim N_3(\beta, \sigma^2 (^tXX)^{-1})$$

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{30})$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{10^6})$$

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{10^4})$$ estimateurs gaussiens non corrélés donc indépendants

$$\hat{\sigma}^2 = \frac{\left\| P - X\hat{\beta} \right\|^2}{30-3} = \frac{1}{27}\sum_{t=1}^{30}(P_i - \hat{\beta}_0 - \hat{\beta}_1 R_i - \hat{\beta}_2 T_i)^2 = \frac{1}{27}\sum_{t=1}^{30}(P_i - \frac{1}{30}\sum_{k=1}^{30} P_k - 10^{-6} R_i \sum_{k=1}^{30} R_k P_k - 10^{-4} T_i \sum_{k=1}^{30} T_k P_k)^2$$

2)

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, $\frac{\hat{\beta}_2}{\sqrt{\frac{\hat{\sigma}^2}{10^4}}} = \frac{100\hat{\beta}_2}{\sqrt{\hat{\sigma}^2}}$ suit une loi de Student à n-3=27 degrés de liberté,

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

Si p_value $< \alpha$: rejet de H_0 au niveau α (la température a un effet significatif)

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

(l'effet de la température est jugé non significatif).

3)

a)

Les estimations calculées de β_1 et β_2 sont respectivement 0.043 m^2 et -0.004 $W.°C^{-1}$

La p-value du test de Student relatif à β_2 (0.972) étant supérieure à 0.05, l'échantillon disponible ne permet pas le rejet de H_0 au niveau de risque 5%. On conclut donc que l'effet de la température ne peut pas être considéré comme significatif.

Remarque : ce résultat pouvait surprendre car l'effet de la température sur le rendement d'un panneau photovoltaïque est connu par ailleurs, le rendement décroissant quand la température augmente, β_2 devant être négatif. Mais nous disposons ici d'un petit échantillon et le résultat n'apparait pas comme significatif sur ces données.

Concernant le rayonnement, la p-value du test de significativité du paramètre β_1 étant faible (0.00021) H_0 est rejetée au niveau 5%. Le prédicteur rayonnement a donc un impact jugé significatif sur le productible photovoltaïque. A température constante, quand le rayonnement augmente de 1 $W.m^{-2}$ la puissance photovoltaïque augmente de 0.043 W.

b)

La valeur manquante est la valeur prise sur les données par la statistique du test de Student relatif au terme constant β_0. On sait qu'elle vaut :

$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} = \frac{80.5}{2} = 40.25$$

c)

On lit sous le tableau fourni : $\hat{\sigma}^2 = 10.95^2$

d)

Ce modèle explique 68.65% de la variance du prédictand. Il modélise cette part de la variabilité de la production photovoltaïque.

4)

On introduit un prédicteur qualitatif (facteur) CAPT disposant de 3 modalités codées par les noms des capteurs utilisés a, b et c. Sur les 30 mesures, nous disposons d'après le tableau de 10 mesures par capteur. La contrainte imposée pour identifier le modèle statistique est de prendre le capteur a comme référence.

Pour rappel, le modèle est alors reconsidéré ainsi, pour des raisons techniques (matrice ^tZZ non inversible sans contrainte) :

$P_i = (\beta_0 + \alpha_a) + \beta_1 R_i + \beta_2 T_i + (\alpha_{capt} - \alpha_a) + e_i$, soit :

$$P_i = \beta_0' + \beta_1 R_i + \beta_2 T_i + \alpha_{capt}' + e_i$$

Avec $\beta_0' = \beta_0 + \alpha_a$ et $\alpha_{capt}' = \alpha_{capt} - \alpha_a$, on a alors :

$\alpha_a' = \alpha_a - \alpha_a = 0$, $\alpha_b' = \alpha_b - \alpha_a$ et $\alpha_c' = \alpha_c - \alpha_a$

Les paramètres alors estimés ne sont pas les effets absolus de chaque modalité mais leurs effets différentiels par rapport à la modalité a. Toutes les estimations relatives au facteur CAPT seront donc à interpréter par rapport au capteur a.

Avec cette contrainte nous obtenons alors pour Z et β (attention à la cohérence entre Z, β et l'organisation des données dans votre archive) :

$$\beta = \begin{pmatrix} \beta_0' \\ \beta_1 \\ \beta_2 \\ \alpha_b' \\ \alpha_c' \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & R_1 & T_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & R_{10} & T_{10} & 0 & 0 \\ 1 & R_{11} & T_{11} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & R_{20} & T_{20} & 1 & 0 \\ 1 & R_{21} & T_{21} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & R_{30} & T_{30} & 0 & 1 \end{pmatrix}$$

La dimension du modèle 2, dimension du vecteur β, est donc q=5.

Les paramètres estimés α_b' et α_c' représentent les écarts moyens de production entre ces 2 capteurs et la référence que constitue le capteur a.

On cherche à mettre en évidence une éventuelle différence de réponse entre les 3 panneaux photovoltaïques, soumis aux mêmes données météorologiques. L'analyse des sorties du logiciel R concernant le modèle 2 permet effectivement de noter une différence de comportement :

La p-value du test de Student relatif au paramètre α_b' est de $0.0462 < 0.05$ → rejet de H_0 au niveau 5%
L'écart estimé à 3.21W est jugé significatif, la réponse du capteur b est différente de celle du capteur a.

La p-value du test de Student relatif au paramètre α_c' est de $0.9432 > 0.05$ → non rejet de H_0.
L'écart estimé à 0.11W est jugé non significatif, les capteurs a et c ont des réponses proches.

Deux des panneaux (a et c) ont des réponses similaires alors que le capteur b présente par rapport à la référence prise un biais estimé à +3.11W.

Exercice 2 :

Partie I :

1) $PO = X\beta + e$

Modèle linéaire gaussien, linéarité selon les paramètres β_k, de dimension $q = 4 = \dim(\beta)$.
Les erreurs sont supposées normales, centrées, indépendantes et de variance constante (homoscédasticité).

$$X = \begin{pmatrix} 1 & T_1 & FF_1 & T_1.FF_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_{80} & FF_{80} & T_{80}.FF_{80} \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

2)

$$^tXX = \begin{pmatrix} 80 & 0 & 0 & 0 \\ 0 & 80 & 0 & 0 \\ 0 & 0 & 80 & -40 \\ 0 & 0 & -40 & 40 \end{pmatrix} \quad \text{, et donc :} \quad (^tXX)^{-1} = \frac{1}{80}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix}$$

3)

On obtient alors pour les 4 estimateurs : $\hat{\beta} = (^tXX)^{-1}\,^tXPO = \frac{1}{80}\begin{pmatrix} \sum_{i=1}^{80} PO_i \\ \sum_{i=1}^{80} T_i.PO_i \\ 2\sum_{i=1}^{80} FF_i.PO_i(1+T_i) \\ 2\sum_{i=1}^{80} FF_i.PO_i(1+2T_i) \end{pmatrix}$

$\hat{\beta} \sim N_4(\beta, \sigma^2(^tXX)^{-1})$

$COV[\hat{\beta}_2, \hat{\beta}_3] = \sigma^2(^tTT)^{-1}_{34} = \frac{\sigma^2}{40} \neq 0$, ces 2 estimateurs ne sont donc pas indépendants.

4)

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, $\frac{\hat{\beta}_2}{\sqrt{\frac{\hat{\sigma}^2}{40}}}$ suit une loi de Student à 80-4=76 degrés de liberté,

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

Si p_value $< \alpha$: rejet de H_0 au niveau α (la force du vent a un effet significatif)

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

(l'effet de la force du vent est jugé non significatif).

5)

a)

L'estimation de β_3 vaut -0.01

La p-value du test de significativité étant supérieure à 0.05 (0.54774), H_0 ne peut être rejetée au niveau 5%. On conclut donc que le terme d'interaction est sans intérêt.

b)

R^2=0.6, le modèle 1 explique donc 60% de la variabilité de l'indice PO.

c)

$$\hat{\sigma}^2 = \frac{\left\| PO - X\hat{\beta} \right\|^2}{n-q} = \frac{1}{76}\sum_{i=1}^{80}(PO_i - \hat{\beta}_0 - \hat{\beta}_1 T_i - \hat{\beta}_2 FF_i - \hat{\beta}_3 T_i.FF_i)^2 = 0.15^2$$

d)

La variable T ayant été normée avant l'estimation du modèle et l'indice PO étant sans unité, β_1 est également sans unité. On lit dans le tableau :

$$\hat{\sigma}_1 = 0.016 = \sqrt{\hat{\sigma}^2 ({}^tXX)^{-1}_{22}}$$

L'estimation de β_1 vaut 0.15

La p-value du test de significativité étant très faible, H_0 est rejetée au niveau 5%.

La température a un impact significatif dans notre modélisation, l'indice PO augmentant avec le réchauffement (β_1>0). A FF constant, et en négligeant l'impact du terme d'interaction jugé non significatif, une augmentation de T d'une unité induira une augmentation de l'indice PO de 0.15

e)

Après analyse des p-values des tests, on conserve un modèle avec un terme constant et les 2 prédicteurs T et FF (car p-values<0.05).

Partie II :

1)

La modalité Est du facteur DD est prise comme référence. Les analyses seront donc menées par rapport à un vent d'Est. Cette contrainte induit de reconsidérer le modèle ainsi :

$PO_i = (\beta_0 + \theta_E) + \beta_1 T_i + \beta_2 FF_i + (\theta_i - \theta_E) + e_i$, soit :

$$PO_i = \beta_0' + \beta_1 T_i + \beta_2 FF_i + \theta_i' + e_i$$

Avec $\beta_0' = \beta_0 + \theta_E$ et $\theta_i' = \theta_i - \theta_E$, on a alors :

$\theta_E' = 0$, $\theta_N' = \theta_N - \theta_E$, $\theta_O' = \theta_O - \theta_E$ et $\theta_S' = \theta_S - \theta_E$

Les estimations concernent donc les effets différentiels des différents secteurs de vent par rapport à un vent d'Est.

Avec cette contrainte nous obtenons alors pour Z et β :

$$\beta = \begin{pmatrix} \beta_0' \\ \beta_1 \\ \beta_2 \\ \theta_N' \\ \theta_O' \\ \theta_S' \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & T_1 & FF_1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_{20} & FF_{20} & 1 & 0 & 0 \\ 1 & T_{21} & FF_{21} & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_{40} & FF_{40} & 0 & 1 & 0 \\ 1 & T_{41} & FF_{41} & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_{60} & FF_{60} & 0 & 0 & 1 \\ 1 & T_{61} & FF_{61} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_{80} & FF_{80} & 0 & 0 & 0 \end{pmatrix}$$

La dimension du modèle 2, dimension du vecteur β, est donc q=6.

Les paramètres estimés θ_N', θ_O' et θ_S' représentent les effets moyens sur l'indice PO de vents de secteurs Nord, Ouest et Sud par rapport à l'effet moyen d'un vent d'Est.

2)

a)
L'analyse des 3 dernières p-values du tableau mène aux conclusions suivantes :
Au risque 5%, des vents de secteurs Ouest et Sud ont le même effet moyen sur l'indice PO qu'un vent de secteur Est.
Seule la direction Nord entraîne un effet moyen différent de celui induit par une direction Est, l'écart jugé significatif étant de +0.18, on peut dire qu'un vent de Nord favorise en moyenne des valeurs d'indice plus fortes qu'en situation de vent d'Est, Ouest ou Sud.
L'effet différentiel d'une des modalités du facteur DD étant jugé significatif (modalité Nord), il faut conserver le prédicteur DD car le modèle sera alors plus précis selon la direction du vent.

b)
L'effet différentiel de la modalité Ouest par rapport à la modalité Est est jugé non significatif (p-value>0.05), ce qui ne veut pas dire que cette modalité n'a aucun impact sur le prédictand.
On considère que $\theta_O' = 0 = \theta_O - \theta_E$, les modalités Ouest et Est ont donc le même effet moyen.

c)
Les R^2 des modèles 1 et 2 valent respectivement 0.6 et 0.74 ce qui laisserait penser que le modèle 2 est plus performant, mais les modèles n'ont pas la même dimension.
Le R^2 augmentant 'mécaniquement' avec la dimension d'un modèle, il faudrait effectuer un test (de Fisher-Snédécor) pour conclure rigoureusement sur la significativité de l'écart constaté.

d)
Les modèles testés sont des modèles linéaires gaussiens, le prédictand étant positif, la loi normale, à support non borné, n'est pas vraiment adaptée. En effet rien n'empêchera ces modèles, selon les valeurs prises par les prédicteurs, de parfois prévoir des valeurs négatives de l'indice PO.
De plus, on cherche ici à prévoir un dépassement de seuil de concentration d'ozone correspondant à de fortes valeurs d'indice (PO >1), la régression logistique serait beaucoup plus adaptée et performante dans ce cas (→ voir suite du cours sur les modèles linéaires généralisés).