Chapitre II : Tests de signification et validation du modèle

II.1. Rappels sur le calcul matriciel

Soit A et B deux matrices telles que

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} et\ B=\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

Non commutatif : **AB** n'est pas égal à **BA** en général.

Somme de matrices

$$C=A+B=\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{pmatrix}$$ qui peut s''écrire $c_{ij}=a_{ij}+b_{ij}$

Produit d'une matrice par un réel

Soit $\mu \in \Re$ on a

$$C=\mu A=\begin{pmatrix} \mu a_{11} & \mu a_{12} & \mu a_{13} \\ \mu a_{21} & \mu a_{22} & \mu a_{23} \\ \mu a_{31} & \mu a_{32} & \mu a_{33} \end{pmatrix}$$ qui peut s''écrire $c_{ij}=\mu a_{ij}$

Produit de matrices

C=A.B qui s''écrit $c_{ij}=\sum a_{ik}\, b_{kj}$

Attention A.B≠B.A.

Matrice identité

On définit la matrice identité, notée Id telle que :

$$A.Id=Id.A=A \text{ et } \quad Id=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Une matrice diagonale si $a_{ij}=0$, pour $i \neq j$.

Transposition

Définition : On appelle transposée de la matrice $A=(a_{ij})$ la matrice $B=(b_{ij})$ de terme général $b_{ij}=a_{ji}$. On la note ${}^tA$.

Exemple

$$A=\begin{pmatrix}1&2&3\\4&5&6\\7&8&9\end{pmatrix} \quad A^t=\begin{pmatrix}1&4&7\\2&5&8\\3&6&9\end{pmatrix}$$

Une matrice A réelle **est symétrique** $A=A^t$

II.2. Analyse statistique des résultats

II.2.1. Degrés de liberté

Soit *n* réponses mesurées indépendamment les unes des autres. Il n'existe pas de relation mathématique entre elles. Les *n* écarts à la moyenne correspondants ne sont pas indépendants. En effet, il existe une relation mathématique entre ces écarts. Quand on en connaît *n* – 1, on peut calculer le dernier avec la relation mathématique. Par exemple, reprenons les quatre écarts à la moyenne de l'exemple (voir l'écart-type). Les trois premiers écarts sont : –0,4 +1,1 –1,1

Le quatrième écart s'obtient facilement puisque la somme des écarts est toujours égale à 0 :

quatrième écart : – 0,4 + 1,1 – 1,1 = 0

quatrième écart = 0,4. Il n'y a donc que *n* – 1 écarts indépendants. On dit que la série des *n* écarts à la moyenne possède *n* – 1 *degrés de liberté* (ou ddl). Le nombre de degrés de liberté est important car il intervient dans de nombreuses formules de statistique

II.2.2. Test de Fisher

L'objectif de l'analyse globale des résultats est de définir la qualité descriptive du modèle au moyen d'un tableau d'analyse de la variance, Analysis Of Variance (ANOVA). Pour ce faire, plusieurs grandeurs doivent être préalablement définies. Soit SCT la somme des carrés totale, c'est-à-dire la somme des carrés des écarts entre les mesures de la réponse et leur moyenne :

$$SCT=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \qquad (II.1)$$

Cette somme peut être décomposée en deux sommes, SCM, la somme des carrés due à la régression ou variation expliquée par le modèle et SCE, la somme des carrés des résidus ou variation inexpliquée par le modèle :

$$SCT = SCE + SCM \qquad (II.2)$$

SCM est la somme des carrés des erreurs entre les réponses estimées et la moyenne des réponses mesurées :

$$SCM=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \qquad (II.3)$$

SCE est la somme du carré des écarts entre les réponses mesurées et estimées :

$$SCE= \sum_{i=1}^{n}\left(y_i-\hat{y}_i \right)^2 \quad (II.4)$$

On effectue alors le test de Fisher. F_{cal} est une valeur calculée d'une valeur F de Fisher, à (p-1) et (n - p) degrés de liberté. On calcule le ratio :

$$F_{cal}= \frac{SCM/p-1}{SCE/n-1} \quad (II.5)$$

En pratique, le modèle utilisé contient un terme constant a_o, correspondant à la moyenne des réponses mesurées. Cette composante n'étant d'aucun intérêt dans l'analyse de la variance, elle est supprimée et donc on prend (p-1) degré de liberté pour le modèle de régression.

Pour réunir ces informations, on utilise le tableau de la variance suivant :

Tableau II.1 : Analyse de la variance (ANOVA).

Source de variation	**ddl**	**Variation**	**Carré moyen**	**Fisher**
Régression	p-1	$\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y} \right)^2$	$\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y} \right)^2 / p-1$	$\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y} \right)^2/p-1}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i \right)^2/n-p}$
Résiduelle	n-p	$\sum_{i=1}^{n}\left(y_i-\hat{y}_i \right)^2$	$\sum_{i=1}^{n}\left(y_i-\hat{y}_i \right)^2/n-p$	
Totale	n-1	$\sum_{i=1}^{n}\left(y_i-\overline{y} \right)^2$		

On note F_{crit} (p–1, n– p) la valeur critique au seuil α d'une loi de Fisher à (p–1) et (n–p) degrés de liberté avec une probabilité α si : $F_{cal}>F_{crit}$ (p-1 ; n-p)

II.2.3. Analyse statistique des coefficients (Test de Student)

Les différents paramètres du modèle peuvent aussi être analysés statistiquement. L'hypothèse nulle (H_0) est alors étudiée pour chacun des coefficients, selon laquelle ceux-ci sont nuls. Pour ce faire, la statistique t_{cal} qui dépend de l'estimation de l'écart type de a_i, $\sigma(a_i)$ est alors calculée :

$$t_{cal}= \frac{a_i}{\sigma(a_i)} \quad (II.6)$$

$\sigma(a_i)$: Ecart type des coefficients

$$\sigma(a_i) = \sqrt{\frac{1}{n}\left(\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i \right)^2}{n-p} \right)} \quad (II.7)$$

Pour réaliser ce test au seuil α, il faut comparer la valeur de t de Student avec la valeur critique d'un Student à (n–p) degrés de liberté.

On utilise une table de Student à (n–p) degré de liberté, α étant choisi, on lit dans cette table de Student la valeur t critique (α, n–p). On rejette H_o lorsque $t_{cal} > t_{crit}$.

Si l'hypothèse H_0 est acceptée, cela veut dire que l'effet en question n'est pas, au risque de 0,05, significativement différent de «0» et donc que la variable qui lui est associée n'a pas d'influence sur la réponse.

II.2.4. Coefficient de détermination (R^2)

Le coefficient de détermination R^2 est à la fois la fraction des variations de la réponse expliquée par le modèle et un indice de la qualité de la régression :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCE}{SCT} \qquad \text{(II.8)}$$

$R^2 = 1$, indique un ajustement parfait, par contre un R^2 qui vaut 0 indique l'absence de relation entre la variable dépendante et la variable explicative. Cependant, dans le contexte de la régression multiple, cela pose le problème de la paramétrisation du modèle. Plus l'on ajoute de variables explicatives, plus le R^2 augmente. Pour éviter ce phénomène, on calcule le coefficient de détermination ajusté :

$$R^2_{ajusté} = 1 - \frac{\frac{SCE}{n-p}}{\frac{SCT}{n-1}} \qquad \text{(II.9)}$$

La qualité du modèle sera donc d'autant meilleure que $R^2_{ajusté}$ sera proche de 1.