Le modèle linéaire général : régression et ANCOVA

J.N. Bacro
master ESDB, Université Montpellier

2023/2024

Le modèle statistique ?

Soit $(y_j, z_j)_{1 \leq j \leq N}$ les réalisations d'une variable réponse Y et d'une co-variable Z. Le modèle considéré est le suivant :

$$y_j = b + az_j + e_j \, 1 \leq j \leq N$$

où $(e_j)_j$ désignent les résidus du modèle, supposés vérifier les postulats du modèle linéaire.

En d'autres termes, sous les postulats,

$$Y_i \mid z_i \sim N(b + az_i, \sigma^2), 1 \leq i \leq N$$

Ex :
Ecrire le modèle sous la forme $\mathbf{Y} = X\theta + \mathbf{E}$ et déduire les expressions générales des estimateurs de b et a.
Dans quel cas pourrait-on avoir des problèmes pour l'estimation ?

Après calculs,

$$\begin{aligned}\hat{b} &= \frac{\sum z_i^2 \sum y_i - \sum z_i \sum z_i y_i}{N\sum(z_i-\bar{z})^2} \\ \hat{a} &= \frac{N\sum z_i y_i - \sum z_i \sum y_i}{N\sum(z_i-\bar{z})^2} = \frac{\sum(z_i-\bar{z})(y_i-\bar{y})}{\sum(z_i-\bar{z})^2}\end{aligned}$$

Rmq :

dans le système ${}^tXX\theta = {}^tX\mathbf{Y}$, on a l'équation $\sum_i y_i = Nb + a\sum z_i$, ce qui permet de déduire

$$\hat{b} = \bar{y} - \hat{a}\bar{z}$$

$\rightsquigarrow$ *la droite de régression passe par le point* $(\bar{z}, \bar{y})$.

On a :

1. $\mathbb{E}(\hat{\theta}) = \theta$;
2. $\mathbb{V}(\theta) = ({}^tXX)^{-1}\mathbb{V}(\mathbf{Y}) = \sigma^2({}^tXX)^{-1}$, d'où

$$\begin{aligned}
\mathbb{V}(\hat{b}) &= \frac{\sum z_i^2}{N\sum(z_i-\bar{z})^2}\sigma^2 = \left(\frac{1}{N}+\frac{\bar{z}^2}{\sum(z_i-\bar{z})^2}\right)^2\sigma^2 \\
\mathbb{V}(\hat{a}) &= \frac{1}{\sum(z_i-\bar{z})^2}\sigma^2 \\
\mathbb{COV}(\hat{a},\hat{b}) &= -\frac{\sum z_i}{N\sum(z_i-\bar{z})^2}\sigma^2
\end{aligned}$$

Rmq : Les paramètres a et b sont corrélés ... *(surprenant ?)*.

En utilisant

$$\hat{\sigma}^2 = \frac{\sum(y_i-\hat{a}z_i-\hat{b})^2}{N-2}$$

on déduit les expressions des estimateurs $\widehat{\mathbb{V}(\hat{a})}$, $\widehat{\mathbb{V}(\hat{b})}$ et $\widehat{\mathbb{COV}(\hat{a},\hat{b})}$.

Remarques :

1. Dans les estimations précédentes, le terme $\sum(z_i - \bar{z})^2$ apparaît au dénominateur . . . quelle interprétation pratique peut-on donner de ce résultat ?
2. si les $(z_i)_{1 \leq i \leq N}$ vérifient $\sum z_i = 0$ que vaut $\mathbb{V}(\hat{b})$, $\mathbb{V}(\hat{b})$ et $\mathbb{COV}(\hat{a}, \hat{b})$? quelle interprétation ?
3. intervalles de confiance pour les paramètres ?
   on doit distinguer les IC pour a et b pris de façon individuelle et la région de confiance associée aux deux paramètres.

   $\hat{\theta} = {}^t(XX)^{-1}\,{}^tX\mathbf{Y}$ et sous les postulats du modèle linéaire, on a $\hat{\theta} \sim N(\theta, \mathbb{V}(\hat{\theta}))$, avec $\mathbb{V}(\hat{\theta})$ matrice de var-covar de ${}^t(\hat{a}, \hat{b})$.

   On en déduit par exemple que $\dfrac{\hat{a} - a}{s(\hat{a})} \sim T(n-2)$, d'où la construction d'un IC de confiance $(1 - \alpha)$.

Table d'analyse de variance d'une régression simple

En utilisant une approche géométrique, on peut facilement voir que

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

où $\hat{y}_i = \hat{b} + \hat{a} z_i$, valeur prédite par le modèle.

Soit M_0 le modèle de régression constant correspondant à $\mathbb{E}(y_i) = b$ et M_1 le modèle de régression simple. L'équation précédente peut se réécrire :

$$SCR_{M_0} = SCR_{M_1} + SCM$$

avec SCM : somme des carrés due au modèle de régression M_1.

Comparer les modèles emboités M_0 et M_1 se fait via l'approche déjà vue en Anova et donne lieu à la table d'analyse de variance de la régression :

Table d'analyse de variance d'une régression simple

Source variation	Degrés de liberté	Somme des carrés	Carrés moyens	F
Modèle	1	$\sum_i (\hat{y}_i - \bar{y})^2$	$\frac{\sum_i (\hat{y}_i - \bar{y})^2}{1} \equiv V_A$	$\frac{V_A}{V_R}$
Résiduelle	$N-2$	$\sum_i (y_i - \hat{y}_i)^2$	$\frac{\sum_i (y_i - \hat{y}_i)^2}{(N-2)} \equiv V_R$	
Totale (cor)	$N-1$	SCT_c		

Rmq : $\sum_i (\hat{y}_i - \bar{y})^2 = \hat{a}^2 \sum_i (z_i - \bar{z})^2$;

on voit que $\sum_i (\hat{y}_i - \bar{y})^2$ est nulle ssi la droite de régression est horizontale, i.e. $\hat{a} = 0$; dans le cas contraire, $\sum_i (\hat{y}_i - \bar{y})^2$ est proportionnelle à $\hat{a}^2$

$\rightsquigarrow$ la valeur de $\sum_i (\hat{y}_i - \bar{y})^2$ donne donc une indication de la force de la liaison entre la réponse et la co-variable.

Tests sur les paramètres d'une régression simple

1. la comparaison des modèles M_0 et M_1 précédents revient à tester H_0 $a = 0$ contre H_1 $a \neq 0$. Le test est un test $F(1, N-2)$
2. de façon équivalente, on peut utiliser un test de Student : sous H_0, $\frac{|\hat{a}|}{s_{\hat{a}}} \sim T(N-2)$; dans ce cas, pour mémoire $1/s^2_{\hat{a}} = (N-2)\frac{\hat{a}^2 \sum_i (z_i - \bar{z})^2}{SCR_{M_1}}$
3. test de H_0 $b = 0$ contre H_1 $b \neq 0$: approche Student (ou comparaison modèles).
4. pour tester H_0 $a = a_0$ et $b = b_0$ contre H_1 "au moins une diffère", quelle approche possible ? Statistique de test ?

Tests sur les paramètres d'une régression simple

1. la comparaison des modèles M_0 et M_1 précédents revient à tester H_0 $a = 0$ contre H_1 $a \neq 0$. Le test est un test $F(1, N-2)$
2. de façon équivalente, on peut utiliser un test de Student : sous H_0, $\frac{|\hat{a}|}{s_{\hat{a}}} \sim T(N-2)$; dans ce cas, pour mémoire $1/s^2_{\hat{a}} = (N-2)\frac{\hat{a}^2 \sum_i (z_i - \bar{z})^2}{SCR_{M_1}}$
3. test de H_0 $b = 0$ contre H_1 $b \neq 0$: approche Student (ou comparaison modèles).
4. pour tester H_0 $a = a_0$ et $b = b_0$ contre H_1 "au moins un diffère", quelle approche possible ? Statistique de test ?

Modèles emboités : $M_0 : \mathbb{E}(Y|z) = b_0 + a_0 z$; $M_1 : \mathbb{E}(Y|z) = b + az$

Statistique de test : $F = \frac{(SCR_{M_0} - SCR_{M_1})/2}{SCR_{M_1}/(N-2)}$

Qualité d'une régression linéaire simple

1. Coefficient de détermination (% de variabilité expliquée par le modèle)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SCR_{M_1}}{SCR_{M_0}}$$

   On voit que $(1 - R^2)SCR_{M_0} = SCR_{M_1}$ $\rightsquigarrow$ quantifier la réduction de la SCR_{M_0} quand on passe à M_1.

2. Coefficient de corrélation linéaire.
   Si l'on considère les couples de v.a. (Y_i, Z_i), $1 \leq i \leq N$, on peut s'intéresser à la dépendance en moyenne entre les valeurs de Z et Y, **sans préjuger a priori d'une variable réponse, d'une co-variable et d'un lien linéaire.**

   On se pose simplement la question : lorsqu'une variable augmente ou diminue, que peut-on dire du comportement en moyenne de l'autre ?

   $\rightsquigarrow$ *coefficient de corrélation linéaire.*

On définit :

$$\rho = \frac{\mathrm{COV}(Y, Z)}{\sigma_Z \sigma_Y}$$

On a les propriétés suivantes :

1. $-1 \leq \rho \leq 1$;
2. $\rho = 0$: absence de corrélation linéaire ;
3. $|\rho| = 1$: liaison linéaire exacte, i.e. $y_i = b + az_i$;
4. **nombreux pièges pour l'interprétation !** en particulier,

   $\rho \approx 0 \Leftrightarrow$ absence de corrélation linéaire mais $\not\Rightarrow$ absence de relation !
   $|\rho| \approx 1 \Leftrightarrow$ corrélation linéaire mais $\not\Rightarrow$ dépendance linéaire !
5. estimation : $\hat{\rho} = \dfrac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sqrt{\sum_i (z_i - \bar{z})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$
6. **Lien R^2 et ρ ?**
   $\hookrightarrow$ pour la régression linéaire **simple**, $R = |\rho|$.

Pour la régression linéaire simple,
toujours faire un dessin du nuage (z_i, y_i) !

Ayant ajusté un modèle de régression simple de Y sur Z, on veut, pour une valeur $z = z_0$ fixée, prédire la valeur associée de la réponse.

Il faut distinguer 2 types de prédiction :

1) prédire une *réalisation moyenne* de la réponse pour $z = z_0$

   $\rightsquigarrow$ modèle $\mathbb{E}(Y) = b + az$;

2) prédire une *réalisation particulière* de la réponse pour $z = z_0$

   $\rightsquigarrow$ modèle $y = b + az + e$;

Cas 1) : soit $\mu_z = \mathbb{E}(Y \mid z)$; $\hat{\mu}_{z_0} = \hat{b} + \hat{a}z_0$; $\hat{\mu}_{z_0} \sim N(b + az_0, \mathbb{V}(\hat{\mu}_{z_0}))$;
Comme $\mathbb{V}(\hat{\mu}_{z_0}) = \mathbb{V}(\hat{b}) + z_0^2\mathbb{V}(\hat{a}) + 2z_0\mathbb{COV}(\hat{a}, \hat{b})$, en injectant les résultats précédents, on obtient $\mathbb{V}(\hat{\mu}_{z_0})) = \sigma^2 \left[\frac{1}{N} + \frac{(\bar{z} - z_0)^2}{\sum_i (z_i - \bar{z})^2}\right]$ et

$$\widehat{\mathbb{V}(\hat{\mu}_{z_0})} = \frac{SCR_{M_1}}{N-2}\left[\frac{1}{N} + \frac{(\bar{z} - z_0)^2}{\sum_i (z_i - \bar{z})^2}\right]$$

Rmqs :
Variance $\downarrow$ lorsque $z_0 \to \bar{z}$;
I.C. pour $\mathbb{E}(Y \mid z_0)$: $\hat{b} + \hat{a}z_0 \overset{+}{-} s_{\hat{\mu}_{z_0}} t_{1-\alpha/2}(N-2)$

Cas 2) : on considère le modèle $Y_0 = b + az_0 + E_0$;

$\hat{Y}_0 = \hat{b} + \hat{a}z_0 + \hat{E}_0$ ... et $\hat{E}_0 =$?

$$\begin{aligned}
\mathbb{V}(\hat{Y}_0) &= \mathbb{V}(\hat{b} + \hat{a}z_0 + E_0) \\
&= \mathbb{V}(\hat{b} + \hat{a}z_0) + \mathbb{V}(E_0) \\
&= \sigma^2 \left[\frac{1}{N} + \frac{(\bar{z} - z_0)^2}{\sum (z_i - \bar{z})^2} \right] + \sigma^2 \\
&= \sigma^2 \left[\frac{1}{N} + \frac{(\bar{z} - z_0)^2}{\sum (z_i - \bar{z})^2} + 1 \right]
\end{aligned}$$

et à nouveau σ^2 estimée par $\frac{SCR_{M_1}}{N-2}$.

On voit que dans les cas 1) et 2, *même prédiction ponctuelle* mais *précision* différente

$$\boxed{\mathbb{V}(Y_{predit}) = \mathbb{V}(Y_{moyen\ predit}) + \mathbb{V}(E)}$$

Quelle conséquence sur les I.C. ?

Régression linéaire simple : test de linéarité

Les test précédents $H_0 : a = 0$ contre $H_1 : a \neq 0$ ne permettent de juger de la qualité de la régression (i.e. le modèle est-il un "bon" modèle ?). Pour tester l'intérêt d'un modèle *imposant une forme de relation* entre y et z, à savoir $y = b + az$, il faudrait le comparer à un modèle qui ne fixe aucune relation particulière . . . typiquement à $y_{ij} = \mu + \alpha_i + e_{ij}$!

Pour cela il faut plusieurs observations de Y pour un même z_j. *La SCR de ce modèle ne contient que des fluctuations non contrôlées par l'expérimentateur*.

Comparer le modèle *large* $Y_{ij} = \mu_j + e_{ij}$ (J paramètres) contre le modèle *restreint* $y_{ij} = b + az_j + e_{ij}$ (2 paramètres).

Problème : modèles emboités ?

Définition 2 : Un modèle restreint est emboité dans un modèle large s'il existe une matrice A déterministe vérifiant $X_{large} \times A = X_{restreint}$, où X désigne la matrice de design des modèles.

Ex :

- M_0 $y_{ij} = \mu + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu + \alpha_j + e_{ij}$

Emboités ?

- M_0 $y_{ij} = b + ax_j + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu_j + e_{ij}$

Emboités ?

Ex :

- M_0 $y_{ij} = \mu + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu + \alpha_j + e_{ij}$

$$X_{M0} : \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} ; X_{M1} : \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Emboités :

$$X_{M1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = X_{M_0}$$

- M_0 $y_{ij} = b + ax_j + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu_j + e_{ij}$

Emboités ?

Ex :

- M_0 $y_{ij} = \mu + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu + \alpha_j + e_{ij}$

Emboités :

$$X_{M1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = X_{M_0}$$

- M_0 $y_{ij} = b + ax_j + e_{ij}$, $1 \leq i \leq 2$, $1 \leq j \leq 2$ et M_1 $y_{ij} = \mu_j + e_{ij}$

$$X_{M_0} : \begin{pmatrix} 1 & x_1 \\ 1 & x_1 \\ 1 & x_2 \\ 1 & x_2 \end{pmatrix} ; X_{M_1} : \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Emboités :

$$X_{M_1} \begin{pmatrix} 1 & 0 \\ 0 & x_1 \\ 0 & x_2 \end{pmatrix} = X_{M_0}$$

Du fait de l'emboitement, la SCR du modèle restreint est redécomposée en deux sommes de carrés dont l'une est la SCR du nouveau modèle et l'autre va représenter *le défaut d'ajustement :*

$$\underbrace{\sum_{i,j}(y_{ij}-\hat{y}_j)^2}_{SCR_{M_{rest}}} = \underbrace{\sum_{i,j}(y_{ij}-y_{\cdot j})^2}_{SCR_{M_{large}}} + \underbrace{\sum_{i,j}(y_{\cdot j}-\hat{y}_j)^2}_{SC_{Defaut\ Ajust}}$$

On a la table suivante

Source variation	Degrés de liberté	Somme des carrés	Carrés moyens	F
Modèle	1	$\sum_{i,j}(\hat{y}_j - y_{\cdot\cdot})^2$	$\frac{\sum_{i,j}(\hat{y}_j - y_{\cdot\cdot})^2}{1} \equiv V_A$	$\frac{V_A}{V_R}$
Défaut Ajust	J-2	$\sum_{i,j}(y_{\cdot j} - \hat{y}_j)^2$	$\frac{\sum_{i,j}(y_{\cdot j} - \hat{y}_j)^2}{J-2} \equiv V_B$	$\frac{V_B}{V_R}$
Résiduelle	N-J	$\sum_{i,j}(y_{ij} - y_{\cdot j})^2$	$\frac{\sum_{i,j}(y_{ij} - y_{\cdot j})^2}{(N-J)} \equiv V_R$	
Totale (cor)	N-1	SCT_c		

et le test F sur le défaut d'ajustement permet de conclure.

Rmq : ce dernier test n'est rien d'autre que le test des deux modèles emboités ! La seule connaissance des SCR (et de leurs ddl) permet donc de conclure.

On généralise l'approche précédente en introduisant plusieurs régresseurs (co-variables quantitatives) $Z_1, \ldots, Z_p$.

Le modèle s'écrit

$$\mathbb{E}(Y_i) = b + a_1 Z_{i1} + \ldots + a_p Z_{ip}$$

Matriciellement

$$\mathbf{Y} = X\theta + \mathbf{E}$$

X ? θ ?

Estimation ?

si l'on suppose la non-colinéarité de $Z_1, \ldots, Z_p$, on a vu que chercher ${}^t(b, a_1, \ldots, a_p)$ tels que $\sum_i (y_i - b - a_1 z_{i1} - \ldots - a_p z_{ip})^2$ minimale revient à calculer

$$\hat{\theta} = ({}^tXX)^{-1}\,{}^tX\mathbf{Y}$$

et $\hat{\mathbf{Y}} = X\hat{\theta}$.

$\hat{\mathbf{Y}}$ est la projection $\perp$ de $\mathbf{Y}$ sur $\mathcal{P}$. Matrice de projection P :

$$P = X({}^tXX)^{-1}\,{}^tX$$

Reprenant les écritures matricielles on montre :

$$\mathbb{E}(\hat{\theta}) = \theta$$

$$\mathbb{V}(\hat{\theta}) = \sigma^2({}^tXX)^{-1}$$

Loi de $\hat{\theta}$?

sous les postulats : $\hat{\theta} \sim N(\theta, \sigma^2({}^tXX)^{-1})$

Estimation de σ^2 :

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{N-(p+1)} = \frac{\sum_i (y_i - \hat{y}_i)^2}{N-(p+1)}$$

Tests sur les paramètres :

1. tester $H_0 : a_j = 0$ contre $H_1 : a_j \neq 0$ en utilisant $\frac{|\hat{a}_j|}{\sqrt{\mathbb{V}(\hat{a}_j)}} \sim T(N-(p+1))$
2. utiliser des approches par modèles emboités.

   Pour tester $H_0 : a_{q+1} = a_{q+2} = \ldots = a_p = 0$, avec $p > q$, contre H_1 "au moins un non nul", quelle statistique de test ?

Régression linéaire multiple : qualité ?

Les tests permettent de répondre à la question : les régresseurs sont-ils nécessaires ?

Reste la question : les régresseurs utilisés *sont-ils suffisants* pour expliquer correctement la réponse ?

1. Coefficient de détermination R^2 ; problème : R^2 augmente systématiquement avec le nombre de régresseurs !
2. R^2 ajusté : R^2 pénalisé en tenant compte du nombre de paramètres. $1 - R^2_{ajuste} = (1 - R^2)\frac{N-1}{N-(p+1)}$ d'où

$$R^2_{ajuste} = \frac{(N-1)R^2 - p}{N-(p+1)}$$

On a $1 - R^2 = \frac{SCR_M}{SCR_{M_0}}$ et $(1 - R^2)\frac{N-1}{N-(p+1)} = \frac{SCR_M/(N-(p+1))}{SCR_{M_0}/(N-1)}$, i.e. $1 - R^2_{ajuste}$ = rapport de variance. Si on prend en compte des régresseurs "inutiles", l'estimation de la variance résiduelle n'est pas améliorée et reste comparable à celle sans ces régresseurs et de fait $1 - R^2_{ajuste}$ ne bouge pratiquement pas !

Régression linéaire multiple : qualité ?

1. coefficient $C(p)$ de Mallows. L'idée est de chercher pour un p fixé s'il existe un sous-modèle M_q, $q < p$, expliquant la réponse aussi bien que M_p.

   Idée : $\frac{SCR_{M_q}}{N-(q+1)} \approx \frac{SCR_{M_p}}{N-(p+1)}$, ce qui donne

   $$q + 1 \approx 2(q+1) - N + \frac{SCR_{M_q}}{SCR_{M_p}/(N-(p+1))}$$

   En posant $C(q+1) = 2(q+1) - N + \frac{SCR_{M_q}}{SCR_{M_p}/(N-(p+1))}$, on voit que

   1. $C(q+1) > q+1$, M_q est moins bon que M_p : le modèle à $(q+1)$ paramètres est sous-paramétré (il manque des régresseurs)
   2. $C(q+1) < q+1$, avec moins de paramètres M_q fait mieux en terme de variance résiduelle : on a rajouté aux q régresseurs considérés, $p-q$ régresseurs qui n'améliorent pas le modèle ... M_p est sur-paramètré.
   3. $C(q+1) = q+1$: le modèle M_q est équivalent à M_p, mais avec moins de paramètres.

Régression linéaire multiple : sélection automatique

Méthodes automatiques permettant d'enlever ou d'ajouter un régresseur à un modèle en cours de construction.

1. Méthode forward : on part du modèle le plus simple M_0 $\mathbb{E}(Y_i) = \mu$ et on "rentre" dans le modèle de proche en proche (un par un) les régresseurs le plus significatifs. Quand aucun régresseur n'est significatif, arrêt de la procédure.
2. Méthode backward : on part du modèle complet et on enlève de proche en proche un régresseur en prenant celui qui est le moins significatif. Quand tous les régresseurs restants sont significatifs, arrêt de la procédure.
3. Méthode stepwise : mélange des deux procédures précédentes. Chaque fois qu'un régresseur est entré dans le modèle, on regarde si un des régresseurs déjà dans le modèle peut être enlevé.

Régression linéaire multiple : interaction

• Que signifie que deux régresseurs sont en interaction dans un modèle de régression multiple ?

Ex : supposons $z1$ et $z2$ en interaction

$$\mathbb{E}(y_i) = b + a_1 z_{i1} + \mathbf{a_2(z_{i1})} z_{i2}$$

Cadre de la régression : $a_2(z_{i1}) = c + dz_{i1}$

Le modèle devient :

$$\mathbb{E}(y_i) = b + a_1 z_{i1} + a_2 z_{i2} + a_3 z_{i3}$$

avec $z_{i3} = z_{i1} z_{i2}$

$\rightsquigarrow$ *ajouter une nouvelle variable obtenue comme produit des deux variables considérées !*

• Tester une interaction ?

Régression linéaire multiple : diagnostiques

On cherche à repérer des *individus atypiques*, qui souvent peuvent avoir une influence très importance sur le modèle obtenu, des *variables colinéaires*, des problèmes potentiels dans l'ajustement etc.

Trois grandes familles :

1. Résidus partiels : juger de l'intérêt d'un régresseur Z_j quand tous les autres ont été pris en compte (notion de corrélation partielle) ;
2. Influence des individus sur l'ajustement : juger de la robustesse de l'ajustement. On utilise une série d'indicateurs aidant au jugement quant à l'influence des observations. Les indicateurs sont fondés sur différents critères et cherchent à détecter d'éventuelles influences de différentes nature.
    1. détection d'un effet levier : résidu *trop* faible (h_{ii}) ;
    2. détection d'une observation atypique : résidu *trop* grand (*rstudent*, *rstandardise*) ;
    3. détection d'un effet *trop* important d'une observation sur l'ajustement obtenu (*dffits)*, sur les valeurs d'estimation des paramètres du modèle (*dcook* pour le vecteur des paramètres, *dfbetas* pour chacun des paramètres), sur la précison des estimations (*covratio*).
3. Colinéarité : répérer les dépendences *trop* fortes entre variables (*Vif*).

1. effet levier : h_{ii} "poids" de l'observation y_i sur sa propre estimation $\hat{y}_i$; si h_{ii} grand alors observation i est suspecte ; en général, grand : $> 2p/N$
2. résidus standardisés : $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$. Si $|r_i| > 2$, individu suspect ;
3. résidus studentisés : $r_i^* = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$. Si $|r_i^*| > t_{0.975}(N-p-1)$, individu suspect ;
4. distance de Cook : $Cook_i = \frac{1}{p\hat{\sigma}^2}\,{}^t(\hat{\theta}_{(-i)} - \hat{\theta})({}^tXX)(\hat{\theta}_{-(i)} - \hat{\theta}) = \frac{h_{ii}r_i^2}{(p+1)(1-h_{ii})^2\hat{\sigma}^2} = \frac{\sum_j(\hat{y}_i - \hat{y}_i^{-j})^2}{\hat{\sigma}^2(1+p)}$; si $Cook_i > 1$, individu suspect ;
5. écart standardisés : $\text{ddfits} = \frac{\hat{y}_i - \hat{y}_{(-i)}}{s_{(-i)}\sqrt{h_{ii}}} = |r_i^*|\sqrt{\frac{h_{ii}}{1-h_{ii}}}$; si $> 2\frac{\sqrt{p+1}}{\sqrt{N}}$, individu suspect.
6. Dfbetas : $\frac{\hat{\theta}_j - \hat{\theta}_j^{(-i)}}{\hat{\sigma}_{(-i)}\sqrt{({}^tXX)^{-1}_{j+1,j+1}}}$; individu i suspect si pour au moins un j Dfbetas $> 2/\sqrt{N}$.
7. $Covratio_i = \frac{\hat{\sigma}^2_{(-i)}}{\hat{\sigma}^2}$; si écart à 1 $> \frac{3(p+1)}{N}$, individu suspect.
8. $Vif_j = \frac{1}{1-R_j^2}$; si $Vif_j > 10$ alors problème avec la variable j.

dans une régression multiple, le signe du coefficient attaché à un régresseur n'est pas représentatif de la liaison entre les variations de la réponse et du régresseur considéré !

Cela peut s'expliquer intuitivement par le fait que la valeur obtenue pour le coefficient intègre toutes les dépendances de la réponse avec l'ensemble des régresseurs considérés et des phénomènes de compensation peuvent se produire !

Pour caractériser la liaison entre la réponse et l'un des régresseurs de la régression multiple, on utilise la notion de de coefficient de corrélation partielle.

Idée : quantifier la corrélation linéaire entre Y et le régresseur Z_j *après prise en compte des éventuelles dépendances vis à vis des autres régresseurs considérés !*

Principe : soit le modèle $\mathbb{E}(y_i) = b + a_1 z_{i1} + a_2 z_{i2} + a_3 z_{i3}$
Corrélation partielle entre Y et $Z2$?

1. régresser **Y** sur **Z1** et **Z3** ; soit **E1** les résidus associés.
2. régresser **Z2** sur **Z1** et **Z3** ; soit **E2** les résidus associés.
3. Coefficient cherché $\rho_{\mathbf{Y},\mathbf{Z2}|\mathbf{Z1},\mathbf{Z3}} = \rho_{\mathbf{E1},\mathbf{E2}}$

Régression linéaire multiple et variable qualitative

Soit W une variable qualitative à 3 niveaux $0, 1, 2$ réprésentant par exemple des groupes de fumeurs (NF,F,GF).

Comment tenir compte de W dans une approche régression multiple ?
considérer $W = 0, 1$ *ou* 2 *selon le groupe fumeur ?* . . .
$\hookrightarrow \mathbb{E}(y_i) = b + a_1 z_{i1} + a_2 z_{i2} + a_3 w_i$ est -il d'intérêt ? pourquoi ? (écrire alors le modèle concerné)

Soit W une variable qualitative à 3 niveaux $0, 1, 2$ réprésentant par exemple des groupes de fumeurs (NF,F,GF).

Comment tenir compte de W dans une approche régression multiple ?
considérer $W = 0, 1$ *ou* 2 *selon le groupe fumeur ?* . . .
$\hookrightarrow \mathbb{E}(y_i) = b + a_1 z_{i1} + a_2 z_{i2} + a_3 w_i$ est -il d'intérêt ? pourquoi ? (écrire alors le modèle concerné)

Comment s'en sortir ?

$\rightsquigarrow$ introduire 2 variables $W1$ et $W2$ binaires pour représenter W !

$$\mathbb{E}(y_i) = b + a_1 z_{i1} + a_2 z_{i2} + a_3 w1_i + a_4 w2_i$$

Si variable k modalités, introduire $(k-1)$ variables binaires !

Cette approche peut-être utilisée pour comparer des groupes.

Cas de 2 groupes :

Ex : après injection d'un médicament, on considère la variable réponse Y pression sanguine, la variable âge Z et la variable sexe S. La relation entre Y et Z différe-t-elle selon le sexe ? (en d'autres termes, effet de S sur la liaison ?)

Soit avec $s_i = 0$ ou 1 selon le sexe.
Modèle1 : $y_i = b_0 + a_1 z_i + a_2 s_i + e_i$, $1 \leq i \leq N$.

Modèle2 : $y_i = b_0 + a_1 z_i + a_2 s_i + a_3 z_i d_i + e_i$, $1 \leq i \leq N$.

Dans modèle1, quelle interprétation de "a_2 *significativement différent de 0* " ?
Comment le tester ?

Quels tests vous semblent d'intérêt pour Modèle2 ?

Si on considère une variable qualitative avec 3 modalités, que deviennent les modèles ?

Lorsque les colonnes de X sont corrélées, il peut être intéressant de procéder à un changement de variables en introduisant les composantes principales issues d'une ACP du tableau X.

Comme ${}^tXX$ est symétrique, on a ${}^tXX = P\Lambda\,{}^tP$, où P est une matrice de vecteurs propres normalisés, i.e. P orthogonale ${}^tPP = Id$, et $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$, matrice de valeurs propres, $\lambda_1 \geq \ldots \geq \lambda_p$.

On a

$$\mathbf{Y} = X\theta + \mathbf{E} = XP\,{}^tP\theta + \mathbf{E} = X^*\theta^* + \mathbf{E}$$

où X^* correspond aux composantes principales (d'où l'appellation !), $X_i^* = XP_i$ et ${}^tX_i^*X_i^* = \lambda_i$ car

$${}^tX^*X^* = {}^t\,P\,{}^tXXP = {}^tPP\Lambda\,{}^tPP = \Lambda$$

Les nouvelles variables de X^* sont orthogonales et

$$\hat{\theta}^* = \Lambda^{-1}\,{}^tP\,{}^tX\mathbf{Y}$$

C'est un estimateur MCO car $||\mathbf{Y} - X\theta||^2 = ||\mathbf{Y} - XP\,{}^tP\theta||^2 = ||\mathbf{Y} - X^*\theta^*||^2$ et $\mathbb{V}(\hat{\theta}^*) = \sigma^2({}^tXX)^{-1} = \sigma^2\Lambda^{-1}$, i.e. *non-corrélés !*
Rmq : comme les λ_i sont ordonnées, la variance plus précise sur les premières composantes.

Analyse de covariance

Cadre concerné : Comparaison de k groupes ou étude de l'effet d'un facteur sur une réponse *après prise en compte d'éventuelles dépendances linéaires de la réponse vis-à-vis de co-variables continues.*

Exemple typique :
on observe une variable réponse Y et deux co-variables Z et W dont l'une est quantitative et l'autre qualitative. Exemple : Y : volume expiratoire ; Z : sexe ; W : âge.
On cherche à voir si la réponse Y dépend du sexe (Z). Si les patients hommes sont plus agés que les patients femmes, et si l'âge est lié à Y, une étude directe de la dépendance entre Y et Z pourrait être faussée par le rôle que peut jouer l'âge W *(confusion d'effets)*. L'idée est donc de comparer les moyennes de Y selon la variable sexe $Z1$ *après prise en compte d'un éventuel effet de l'âge sur les valeurs de la réponse*.

Comment illustrer graphiquement le problème posé ?

Modèle additif :

$$y_{ij} = \mu + \alpha_i + aw_{ij} + e_{ij}$$

avec $1 \leq i \leq I$, $1 \leq j \leq J$.

Modèle avec interaction entre la co-variable qualitative et la co-variable quantitative ?

Analyse de covariance

Le problème précédent peut être généralisé à un plus garnd nombre de variables sans difficultés particulières.

• **Estimation des paramètres ?** on utilise l'écriture matricielle et les résultats précédemment vus s'appliquent.

• **Tests sur les paramètres ?** les approches modèles emboités s'appliquent ; si nécessité de comparaisons non-emboités, des critères standard tels que l'AIC (Akaïké) fonctionne.

• **Validité du modèle : ?** mêmes approches que celles vues précédemment.