Le modèle linéaire général gaussien

J.N. Bacro
master ESDB, Université Montpellier

2023/2024

Modèle linéaire général : écriture et estimations des paramètres

Tout modèle linéaire (analyse de variance, régression, analyse de covariance) peut s'écrire sous la forme

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

Deux cas à considérer :

1. ${}^tXX$ inversible, i.e. $\text{rang}({}^tXX) = p$, nombre de paramètres, alors

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

2. ${}^tXX$ non-inversible, $\text{rang}({}^tXX) = r < p$;
   r : nombre de paramètres indépendants. On rajoute $p - r$ contraintes, sous la forme

   $B\theta = \mathbf{0}$, B matrice de dimension $(p - r, p)$.

   On a alors : $\left\{ \begin{array}{lcl} {}^tXX\theta & = & {}^tX\mathbf{Y} \\ B\theta & = & \mathbf{0} \end{array} \right. \Rightarrow \left\{ \begin{array}{lcl} {}^tXX\theta + {}^tBB\theta & = & {}^tX\mathbf{Y} \\ {}^tBB\theta & = & \mathbf{0} \end{array} \right.$

Soit $G = \left(\frac{X}{B}\right)$.

Le système précédent s'écrit alors :

$$ {}^tGG\theta = {}^tX\mathbf{Y} $$

avec G inversible (G de rang plein). D'où

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

où $({}^tXX)^{-}$ désigne une notion d'inverse généralisée.

Soit $\hat{\mathbf{Y}} = \mathbb{E}(\mathbf{Y})$ le vecteur des valeurs prédites par le modèle considéré.

$$ \boxed{\hat{\mathbf{Y}} = X\hat{\theta} = X\,({}^tXX)^{-}\,{}^tX\mathbf{Y} \equiv H\mathbf{Y}} $$

avec H : *hat-matrix*

Propriétés statistiques de $\hat{\theta}$?

1. Biais ?

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \mathbb{E}\left(({}^tXX)^- \, {}^tX\mathbf{Y}\right) &&= \mathbb{E}\left(({}^tGG)^{-1} \, {}^tX\mathbf{Y}\right) \\
&= ({}^tGG)^{-1} \, {}^tX\mathbb{E}(\mathbf{Y}) &&= ({}^tGG)^{-1} \, {}^tXX\theta \\
&= ({}^tGG)^{-1} \left[{}^tXX\theta + {}^tBB\theta\right] &&= ({}^tGG)^{-1} \, {}^tGG\theta \\
&= \theta
\end{aligned}$$

$\hat{\theta}$ estimateur sans biais

2. Variance ?

Si A matrice déterministe, on a $\mathbb{V}(A\mathbf{Y}) = A\mathbb{V}(\mathbf{Y}) \, {}^tA$.

   1. si ${}^tXX$ inversible :

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

   2. si ${}^tXX$ non-inversible :

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

Propriétés de $\hat{\theta}$

Théorème de Gauss-Markov :

L'estimateur des moindres carrés $\hat{\theta}$ est de variance minimale parmi les estimateurs *linéaires* sans biais de θ.

Eléments de preuve :
On suppose ${}^tXX$ inversible, i.e. $\hat{\theta} = ({}^tXX)^{-1}\,{}^tX\mathbf{Y}$.
Soit $\hat{\theta}^* = C\mathbf{Y}$ autre estimateur linéaire sans biais de θ.

$$\theta = \mathbb{E}(\hat{\theta}^*) = C\mathbb{E}\mathbf{Y} = CX\theta$$

donc $CX = Id$.

Soit $M = C - ({}^tXX)^{-1}\,{}^tX \equiv C - A$.

On a alors $M\,{}^tA = A\,{}^tM = \mathbf{0}$. On en déduit :

$$\mathbb{V}(\hat{\theta}^*) = \sigma^2 C\,{}^tC = \sigma^2 M\,{}^tM + \sigma^2({}^tXX)^{-1} = \sigma^2 M\,{}^tM + \mathbb{V}(\hat{\theta})$$

Loi de $\hat{\theta}$?

En s'appuyant sur le modèle et ses postulats, $\mathbf{Y} \sim N(X\theta, \sigma^2 Id)$. D'où

$$\boxed{\hat{\theta} \sim N(\theta, \mathbb{V}(\hat{\theta}))}$$

Estimation de σ^2

Théorème de Cochran :

Si $\mathbf{U} \sim N(\mathbf{0}, Id)$, A matrice non-aléatoire telle que $A^2 = A$, alors

$$^t\mathbf{U}A\mathbf{U} \sim \chi^2(m)$$

où $m = rang(A)$.

On a

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = (Id - P)\mathbf{Y} = (Id - P)(X\theta + \mathbf{E}) = (Id - P)\mathbf{E}$$

où P matrice de projection sur $\mathcal{P}$, espace engendré par les colonnes de X.

$||\hat{\mathbf{E}}||^2 = {}^t\hat{\mathbf{E}}\hat{\mathbf{E}} = {}^t\mathbf{E}(Id - P)^2\mathbf{E} = {}^t\mathbf{E}(Id - P)\mathbf{E}$ et $\mathbf{E} \sim N(\mathbf{0}, \sigma^2 Id)$. D'où

$$\frac{||\hat{\mathbf{E}}||^2}{\sigma^2} = \frac{{}^t\mathbf{E}}{\sigma}(Id - P)\frac{\mathbf{E}}{\sigma} \sim \chi^2(N - r)$$

avec N **nombre total d'observations et** r **nombre de paramètres indépendants du modèle**. Donc $\mathbb{E}(||\hat{\mathbf{E}}||^2) = \sigma^2(N - r)$ et

$$\hat{\sigma^2} = \frac{||\hat{\mathbf{E}}||^2}{N - r} = \frac{\sum_{ij}(y_{ij} - \hat{y}_{ij})^2}{N - r} \equiv \frac{SCR_M}{N - r}$$

estimateur sans biais de σ^2 sous le modèle M.

Tests sur les paramètres du modèle

⋆ **Test sur un paramètre :**

- approche classique

$$H_0 : \theta_i = 0 \text{ contre } H_1 : \theta_i \neq 0$$

On rejette H_0 si

$$\frac{|\widehat{\theta}_i|}{\sigma_{\widehat{\theta}_i}} > t_{1-\alpha/2;N-r} \qquad \widehat{\sigma^2_{\widehat{\theta}_i}} = V(\widehat{\theta})_{ii}$$

$t_{1-\alpha/2;N-r}$ quantile $1-\alpha/2$ de la loi de Student à $N-r$ degrés de liberté.

- approche comparaison de modèles . . . *à suivre !*

⋆ **Test sur un ensemble de paramètres :**

- approche comparaison de modèles . . . *à suivre !*

Mesurer la qualité de la modélisation

Lorsque l'on considère un modèle linéaire M et que l'on souhaite voir s'il est significativement meilleur que le modèle M_0, on fait le test de Fisher via la table d'analyse de variance du modèle. Ce test compare les deux modèles et permet de conclure . . . *mais ne donne aucune info sur la qualité de la modélisation* : **le modèle M représente-t-il correctement les observations ?**

Un **critère subjectif** $\rightsquigarrow$ **le coefficient de détermination R^2 :**

$$R^2 = \frac{\text{Variabilité expliquée par } M}{\text{Variabilité totale}} = \frac{\sum_{i,j}(\hat{y}_{ij} - y_{..})^2}{\sum_{i,j}(y_{ij} - y_{..})^2}$$

$0 \leq R^2 \leq 1$ représente le % de variabilité expliquée par la modèle M ; plus R^2 est proche de 1, meilleure est l'explication . . . mais attention aux pièges d'interprétation !

Cas de l'ANOVA 1 facteur : $\hat{y}_{ij} = y_{.j}$ et

$$R^2 = \frac{\sum_{i,j}(y_{.j} - y_{..})^2}{\sum_{i,j}(y_{ij} - y_{..})^2} = \frac{SCR_{M_0} - SCR_{M_1}}{SCR_{M_0}} = 1 - \frac{SCR_{M_1}}{SCR_{M_0}}$$

Pour généraliser l'approche ANOVA 1 facteur vue précédemment, on se place dans le cadre d'une modélisation statistique particulière, à savoir *le modèle linéaire*.

Modèle ?
traduction mathématique (équation) de l'action ou de l'effet de caractères quantitatifs et/ou qualitatifs (facteurs/ régresseurs) sur une variable aléatoire d'intérêt (la réponse).

Linéaire ?
Espérance de la réponse est une **combinaison linéaire connue de paramètres inconnus**.

Cadre général : à partir d'un nombre fini N de réalisations indépendantes d'une variable réponse Y, on cherche à étudier, caractériser les variations de cette variable selon les valeurs prises par des co-variables potentiellement d'intérêt. On cherche à expliquer, à prédire $\mathbb{E}(Y)$ en fonction de variables (quantitatives et/ou qualitatives) $Z_1, \ldots, Z_p$ dont les réalisations permettent d'exprimer une combinaison linéaire de paramètres inconnus.

Vocabulaire de base :

• *facteur* : représente une cause déterminée suceptible d'être responsable d'éventuelles variations dans les valeurs de la réponse. En général, désigne une variable qualitative.
• *Niveaux d'un facteur* : modalités du facteur concerné.
• Type de facteur :
- *controlé* lorsque les niveaux du facteur sont fixés par l'expérimentateur. L' effet du facteur est fixe.
- *non-controlé* les valeurs sont "subies" par l'expérimentateur. Si observées, elles pourront être prise en compte dans l'étude statistique.
• *traitement* : combinaison des niveaux de chacun des facteurs contrôlés.

Exemples :

Rendement blé selon engrais. Facteur contrôlé : engrais ; éventuellement non-contrôlé : type de sol.

Etude médicale : facteur contrôlé : dose ; facteur contrôlé (ou non !) : âge ; non-contrôlé : rythme cardiaque.

Si 3 types d'engrais et 2 types de sol $\rightsquigarrow$ 6 traitements.

- *expérience factorielle* : expérience où tous les traitements sont appliqués.
- *unités expérimentales* : unités statistiques auxquelles les traitements sont appliqués.
- *répétition* : ensembles des unités recevant un même traitement dans les mêmes conditions.
- *blocs* : regroupements d'unités présentant une caractéristique commune (ex : parcelles selon nature du sol).

Ecriture générale :

$$Y_{tk} = m_t + E_{tk}$$

où
t indice de traitement ;
k indice de répétition du traitement ;
E_{tk} : résidu pour la $k$ème répétition associée au traitement t ;
m_t : traduction de l'effet systématique du traitement t sur les observations ;

l'expression explicite de m_t va dépendre de l'expérience considérée : plusieurs facteurs ? qualitatifs et/ou quantitatifs ? des répétitions ? des blocs ? etc

Postulats du modèle linéaire :

1. $\mathbb{E}(E_{tk}) = 0$;
2. $\mathbb{V}(E_{tk}) = \sigma^2 \equiv Cte$
3. $(E_{tk})_{t,k}$ non-corrélés ;
4. $E_{tk} \sim N(0, \sigma^2)$

Comme $\mathbb{E}(E_{tk}) = 0$, on voit que m_t réprésente l'espérance des réalisations de la réponse pour le traitement t.

La résiduelle E_{tk} englobe toute la variabilité des observations (non expliquée par le modèle) autour de leur espérance, incluant de fait des variabilités diverses telles que variabilité du matériel, facteurs non-controlés, erreurs de mesure, facteurs inconnus, randomistaion ...

L'analyse de variance à 1 facteur a conduit à considérer (M_0) $m_t = \mu$ et (M_1) $m_t = \mu_t$ ou $m_t = \mu + \alpha_t$.

Et si plusieurs facteurs ?
et si un ou des régresseurs ? *(régression simple ou multiple)*
et si mélange des deux cas précédents ? *(ANCOVA)*