Modélisation statistique
Régression multivariée

aurore.lavigne@univ.lille.fr

Partie 2 : Régression linéaire multiple

Le modèle étudié

- Le modèle de régression multiple est une généralisation à plusieurs facteurs (p) du modèle simple. Il s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_1 x_{ip} + \varepsilon_i, i = 1 \ldots, n. \qquad (1)$$

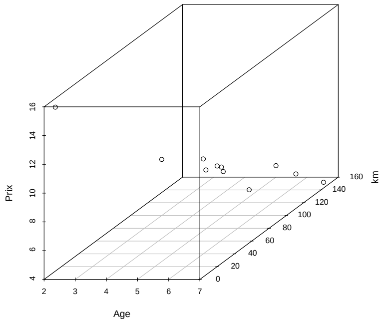
- La terminologie reste la même et on suppose que $n > p + 1$.
  - $\mathbf{Y}$: Variable d'intérêt
  - $\mathbf{x}_1, \ldots, \mathbf{x}_p$: variables explicatives
  - $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$: Vecteur des coefficients ou vecteur des paramètres.

- On est en présence d'un exemple à 2 facteurs auquel on associe le modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

- Ou bien :

$$Y_i = [1\ x_{i1}\ x_{i2}] \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i$$



- La vérification visuelle semble effectivement proche d'un plan.
- Cependant l'interprétation est délicate en présence de 3 facteurs ou plus.

et

$$\mathbf{X} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}}_{\text{matrice } n\times(p+1)} = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \\ 1 & x_i^T \\ \vdots & \\ 1 & x_n^T \end{bmatrix}.$$

Théorème

L'EMCO de $\boldsymbol{\beta}$ est obtenu en résolvant les équations :

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T Y = 0.$$

Il est donné par la formule

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

- Comme par hypothèse $\mathrm{rang}(\mathbf{X}) = p + 1$ c'est à dire $\mathbf{X}$ est de rang maximum, alors $\mathbf{X}^T\mathbf{X}$ est une matrice carré de rang maximum, alors elle est inversible.

Propriétés de l'EMC

Théorème

(i) $\widehat{\boldsymbol{\beta}}$ est un e.s.b. de $\boldsymbol{\beta}$. Sa matrice de variance-covariance est :

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

(ii) $\widehat{\boldsymbol{\beta}}$ est le meilleur estimateur linéaire sans biais de $\boldsymbol{\beta}$ au sens où sa variance est minimale parmi tous les estimateurs linéaires sans biais de $\boldsymbol{\beta}$.

(iii) Si de plus les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, alors $\widehat{\boldsymbol{\beta}}$ correspond à l'estimateur du "Maximum de vraisemblance" de $\boldsymbol{\beta}$ et

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

- On retrouve ainsi la décomposition vue dans le cadre de la régression simple :

$$SCT = SCM + SCR.$$

Estimation de σ^2

Théorème

On considère l'estimateur

$$\widehat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{SCR}{n-p-1}$$

(i) $\widehat{\sigma^2}$ est un e.s.b de σ^2 indépendant de $\widehat{Y}$.

(ii) De plus si les $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ alors :

- $\frac{(n-p-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$.
- Pour tout $j = 0, \ldots, p$, on a $\dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma^2}\,(X^TX)^{-1}_{j+1,j+1}}} \sim \mathcal{T}_{n-p-1}$.

Intervalles de confiance

- Si les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ alors les intervalles de confiance au niveau $1-\alpha$ des β_j sont donnés par :

$$IC(\beta_j) = \left[\widehat{\beta}_j - t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\widehat{\beta}_j}},\ \widehat{\beta}_j - t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\widehat{\beta}_j}}\right],$$

avec

$$\widehat{\sigma_{\widehat{\beta}_j}} = \sqrt{\widehat{\sigma^2}\,(X^TX)^{-1}_{j+1,j+1}}\ :\ \text{erreur standard de } \widehat{\beta}_j.$$

Remarque

Remarque : Si les ε_i ne sont pas supposés gaussiens, alors l'intervalle de confiance précédent est un intervalle de confiance asymptotique, en remplaçant $t_{n-p-1,1-\alpha/2}$ par le fractile $z_{1-\alpha/2}$ de la loi normale.

Tests d'hypothèses : signification d'un coefficient β_j

- Pour $j = 0, \ldots, p$, on souhaite réaliser un test concernant la vraie valeur de β_j.
- Les hypothèses :

$$H_0 : \beta_j = b_j \text{ contre } H_1 : \beta_j \neq b_j, \; b_j \text{ une valeur fixée.}$$

- En particulier $b_j = 0$, on teste l'effet de la variable explicative x_j sur la variable Y
- Réaliser ce test consiste à se demander s'il faut exclure ou non la variable x_j du modèle (on parle de test d'exclusion).
- Si H_0 est rejetée au profit de H_1, on dit que le coefficient β_j est significatif.

Tests sur les paramètres

- Les hypothèses du test :

$$H_0 : \beta_j = b_j \text{ contre } H_1 : \beta_j \neq b_j, \; b_j \text{ une valeur fixée.}$$

- La statistique de test :

$$\frac{\widehat{\beta} - b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} \sim \mathcal{T}_{n-p-1} \text{ sous } H_0.$$

- La règle de décision :
  - si $\left|\frac{\widehat{\beta}-b_j}{\widehat{\sigma_{\widehat{\beta}_j}}}\right| > t_{n-p-1,1-\alpha/2}$, alors je rejette H_0 au profit de H_1.
  - si $\left|\frac{\widehat{\beta}-b_j}{\widehat{\sigma_{\widehat{\beta}_j}}}\right| \leq t_{n-p-1,1-\alpha/2}$, alors je ne rejette pas H_0 au profit de H_1.

Intervalle de confiance d'une combinaison linéaire des coefficients

- On se donne un vecteur $a = (a_0, \ldots, a_p)^T$ et on cherche à estimer $a^T\beta$.
- Exemple : si $a^T = \left(1, x_1^*, \ldots, x_p^*\right)$ où les $x_j^*, j = 1, \ldots, p$ représentent une nouvelle observation des variables explicatives, alors $a^T\beta = \mathbb{E}(Y^*)$.
- On a $a^T\widehat{\beta}$ est un e.s.b de $a^T\beta$, de plus :

$$a^T\widehat{\beta} \sim \mathcal{N}\left(a^T\beta,\ \sigma^2 a^T\left(X^TX\right)^{-1}a\right),$$

ce qui implique :

$$\frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}} \sim \mathcal{T}_{n-p-1}.$$

- Ainsi, un intervalle de confiance au niveau $1-\alpha$ pour $a^T\beta$ est donné par :

$$IC\left(a^T\beta\right) = \left[a^T\widehat{\beta} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}\right].$$

- Dans le cas où $a^T = \left(1, x_1^*, \ldots, x_p^*\right)$ où les x_j^* est une nouvelle observation des variables explicatives, on a :

$$IC\left(\mathbb{E}\left(Y^*\right)\right) = \left[\widehat{Y^*} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}\right].$$

Test d'une contrainte linéaire sur les coefficients

- On souhaite tester une restriction linéaire de la forme :

$$\left\{\begin{array}{l} H_0 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p = b \\ \text{contre} \\ H_1 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p \neq b \end{array}\right.$$

- Exemples :

$$(1)\left\{\begin{array}{l} H_0 : \beta_1 + \beta_2 = 1 \\ \text{contre} \\ H_1 : \beta_1 + \beta_2 \neq 1 \end{array}\right. ; (2)\left\{\begin{array}{l} H_0 : \beta_2 - \beta_3 = 0 \\ \text{contre} \\ H_1 : \beta_2 - \beta_3 \neq 0 \end{array}\right. .$$

- Les hypothèses de test peuvent s'écrire sous forme matricielle :

$$(1)\left\{\begin{array}{l} H_0 : a^T\beta = b \\ \text{contre} \\ H_1 : a^T\beta \neq b \end{array}\right. \quad \text{avec } a^T = (a_0, \ldots, a_p)\,.$$

- Rappel :

$$\frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\widehat{\sigma^2} a^T \left(X^T X\right)^{-1} a}} \sim \mathcal{T}_{n-p-1}.$$

- Ainsi on peut établir la règle de décision suivante :

- si
$$\left|\frac{a^T\widehat{\beta}-b}{\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}}\right| > t_{n-p-1,1-\alpha/2},$$
alors je rejette H_0 au profit de H_1.
- si
$$\left|\frac{a^T\widehat{\beta}-b}{\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}}\right| \leq t_{n-p-1,1-\alpha/2},$$
alors je ne rejette pas H_0 au profit de H_1.

Test de comparaison de deux modèles emboités

- Les hypothèses du test :

$$\begin{cases} H_0 : \beta_{r+1} = \beta_{r+2} = \cdots = \beta_p = 0 \\ \text{contre} \\ H_1 : \exists j \in \{r+1, \cdots, p\} : \beta_j \neq 0 \end{cases}$$

- La statistique de test :

$$\begin{aligned} F = & \frac{[SCR_q - SCR_p]/(p-q)}{SCR_p/n-p-1} = \frac{[SCM_p - SCM_q]/(p-q)}{SCM_p/n-p-1} \\ = & \frac{\left(R_p^2 - R_q^2\right)(n-p-1)}{q(1-R_p^2)} \sim \mathcal{F}_{p-q,n-p-1}, \text{ sous } H_0. \end{aligned}$$

- La règle de décision :
  - si $F > f_{p-q,n-p-1,1-\alpha}$, alors je rejette H_0 au profit de H_1.
  - si $F \leq f_{p-q,n-p-1,1-\alpha}$, alors je ne rejette pas H_0.

Test de signification globale du modèle

- Les hypothèses du test :

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_p = 0 \\ \text{contre} \\ H_1 : \exists j \in \{1, \ldots, p\} \text{ tq } \beta_j \neq 0 \end{cases}$$

- La statistique de test :

$$F = \frac{SCM/p}{SCR/n-p-1} = \frac{R^2(n-p-1)}{p(1-R^2)} \sim \mathcal{F}_{p,n-p-1} \text{ sous } H_0.$$

- La règle de décision :
  - si $F > f_{p,n-p-1,1-\alpha}$, alors je rejette H_0 au profit de H_1.
  - si $F \leq f_{p,n-p-1,1-\alpha}$, alors je ne rejette pas H_0 au profit de H_1.

Critères de comparaison de modèles (emboités ou pas)

- La comparaison de deux modèles se fera avec le R^2_{aj}. Un modèle est donc préférable à un autre si son R^2_{aj} est supérieur à celui de l'autre modèle.
- Il existe d'autres critères du même type permettant de comparer la qualité d'ajustement de 2 modèles relatifs à la même variable dépendante :
  - Le critère d'information d'Akaike (Akaike Information Criterion) :

$$AIC = n \ln\left(\frac{SCR}{n}\right) + 2(p+1).$$

  - Le critère de Bayes de Schwarz (Schwarz Bayesian Information Criterion) :

$$BIC = n \ln\left(\frac{SCR}{n}\right) + (p+1)\ln n.$$

- Contrairement au R^2_{aj}, les critères AIC et BIC sont des fonctions croissantes de SCR.
- Un modèle est donc préférable à un autre au sens du AIC (resp. BIC) si son critère AIC (reps. BIC) est inférieur à celui de l'autre modèle.
- Certains logiciels utilisent une formulation légèrement différente des critères AIC et BIC mais les interprétations restent les mêmes.
- Avec le logiciel **R**, les critères précédents s'obtient avec la fonction `AIC`, qui dépend d'un paramètre k.
- Ce paramètre vaut 2 et $2\ln n$ pour les critères AIC et BIC respectivement.

Intervalle de prévision

- On cherche maintenant un intervalle pour Y^* qui est une variable aléatoire. On a :

$$Y^* = a^T\beta + \varepsilon^* \text{ avec } a^T = \left(1, x_1^*, \ldots, x_p^*\right) \text{ et } \varepsilon^* \sim \mathcal{N}(0, \sigma^2),$$

ε^* indépendant des $\varepsilon_i, i = 1, \ldots, n$.

- Ainsi

$$Y^* \sim \mathcal{N}\left(a^T\beta, \sigma^2\right) \text{ et } \widehat{Y^*} = a^T\widehat{\beta} \sim \mathcal{N}\left(a^T\beta, \sigma^2 a^T\left(X^TX\right)^{-1}a\right),$$

ce qui implique

$$\widehat{e^*} = Y^* - \widehat{Y^*} \sim \mathcal{N}\left(0, \sigma^2\left[1 + a^T\left(X^TX\right)^{-1}a\right]\right).$$

- D'où
$$\frac{Y^* - \widehat{Y^*}}{\sqrt{\widehat{\sigma^2}\left[1 + a^T\left(X^TX\right)^{-1}a\right]}} \sim \mathcal{T}_{n-p-1}.$$
- L'intervalle de prévision est ainsi donnée par :
$$IP\left(Y^*\right)\left[\widehat{Y^*} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}\left[1 + a^T\left(X^TX\right)^{-1}a\right]}\right].$$

Remarque

La quantité
$$a^T\left(X^TX\right)^{-1}a$$
représente le levier de l'observation $x^* = \left(x_1^* \ldots, x_p^*\right)$.