REVUE DE STATISTIQUE APPLIQUÉE

J. ULMO

Problèmes et programmes de régression

Revue de statistique appliquée, tome 19, nº 1 (1971), p. 27-39

<http://www.numdam.org/item?id=RSA_1971__19_1_27_0>

PROBLÈMES ET PROGRAMMES DE RÉGRESSION

J. ULMO

Après un rappel des résultats classiques de l'étude des modèles linéaires, on étudie le problème du choix des variables indépendantes à faire intervenir dans une équation de régression.

On montre que la suppression à tort d'un certain nombre de variables indépendantes conduit en général à des estimateurs biaisés des coefficients de régression relatifs aux variables restantes, puis on passe en revue les principales procédures de choix des variables indépendantes à faire intervenir dans une équation de régression, quand celle-ci est utilisée pour la prévision par "interpolation" (c'est-à-dire quand on se limite au domaine couvert par l'ensemble des valeurs prises par les variables indépendantes dans l'échantillon utilisé).

PLAN

I - INTRODUCTION : RAPPEL DE RESULTATS CLASSIQUES DE L'ETUDE DES MODELES LINEAIRES

1/ Définition d'un modèle de régression linéaire

C'est un modèle du type

$$y = \underline{x}'\underline{\beta} + \varepsilon \tag{1}$$

où $x' = (x^1 \ldots x^p)$(1) représente un ensemble de variables dites indépendantes ou explicatives qu'on traite comme si elles étaient non aléatoires, que ceci soit réalisé ou non, et y est une variable dépendante ou à expliquer, $\underline{\beta} = \left\{ \begin{matrix} \beta_1 \\ \vdots \\ \beta_p \end{matrix} \right.$ est

---

(1) Pour faciliter la compréhension du texte les vecteurs (ou leurs transposés) seront généralement représentés par des lettres soulignées.

le vecteur des coefficients de régression de y sur les $\underline{x}^{\alpha}(\alpha = 1\ldots p)$ et c'est lui qu'on cherche à estimer à partir d'un n échantillon de réalisations $(y_i, \underline{x}'_i)$, $i = 1, 2\ldots n$.

ε est une variable aléatoire de moyenne nulle, en sorte que $E^{x'}(y) = x'\beta$. $E^{x'}(y)$ = espérance de y conditionnée par $\underline{x}'$.

2/ Méthode d'estimation des moindres carrés

Quand on dispose de n réalisations de $(y_i, \underline{x}'_i)$ le modèle correspondant à (1) s'écrit

(2) $y = X\underline{\beta} + \underline{\varepsilon}$ où $\underline{y}$ et $\underline{\varepsilon}$ sont des vecteurs de R^n et $\underset{(n,p)}{X} = ||x_i^{\alpha}|| \left\{ \begin{array}{l} \alpha = 1\ldots p \\ i = 1\ldots n \end{array} \right.$ est une matrice (n, p).

La méthode des moindres corrés consiste à chercher un estimateur $\underline{b}$ de $\underline{\beta}$ tel que $\underline{y}^* = \underline{Xb}$ soit le plus voisin possible de y au sens d'une norme euclidienne de R^n lieu de $\underline{y}$.

Dans la méthode des moindres carrés ordinaire la norme considérée est la norme habituelle définie par le produit scalaire $||y||^2 = \underline{y}'\underline{y} = \sum_{i=1}^{n} y_i^2$.

$\underline{b}$ est donc tel que $\underline{y}^* = X\underline{b}$ et $||\underline{y} - \underline{y}^*||^2 = ||\underline{y} - X\underline{b}||^2 = (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$ soit minimum

Rien n'empêche d'introduire une norme définie par une forme quadratique définie positive de matrice M quelconque, soit $||y||_M^2 = y'My$.

Les seules métriques M utilisées dans la pratique sont celles qui sont caractérisées par une matrice diagonale. $M = D_{w_i}$ de poids w_i. On a alors la méthode des moindres carrés pondérée qui consiste à définir $\underline{b}$ par la condition

$$(\underline{y} - \underline{Xb})'M(\underline{y} - \underline{Xb}) = \sum_{i=1}^{n} w_i (y_i - y_i^*)^2 \text{ minimum avec } y_i^* = \underline{x}'_i\underline{b}$$

En fait on montre que pour conserver les propriétés classiques de la méthode des moindres carrés on doit prendre $M = K\Sigma^{-1}$ où K est une constante si $\Sigma = \text{cov}\,\underline{y} = \text{cov}\,\underline{\varepsilon}$.

On suppose généralement que $\text{cov}\,y = \sigma^2 I_n$ ou $\text{cov}\,y = D_{\sigma_i}^2$ matrice diagonale d'élément σ_i^2.

3 - Interprétation géométrique de la méthode

$y^* = Xb$ est la projection "M orthogonale" de $y \in R^n$ sur le sous-espace linéaire V engendré par les colonnes de X tandis que $\underline{e} = \underline{y} - \underline{y}^*$ vecteur écart résiduel est M orthogonal à V c'est-à-dire est élément de $V^{\perp}$[1] sur lequel il est la projection de y.

On a donc

$$\boxed{y = \underset{\in V}{\underline{y}^*} + \underset{\in V\perp}{\underline{e}}}$$

---

(1) V⊥ (ou V "orthogonal") désigne le sous espace supplémentaire de V dans R^n qui lui est M orthogonal

et $$y'My = y^{*}{'}My^{*} + e'Me \quad \text{car} \quad y^{*}{'}Me = 0$$

soit

(3) $$\boxed{\|y\|^2 = \|y^*\|^2 + \|e\|^2}$$ théorème de Pythagore généralisé.

Dans le cas où M = I, auquel on se bornera par la suite, (3) s'écrit

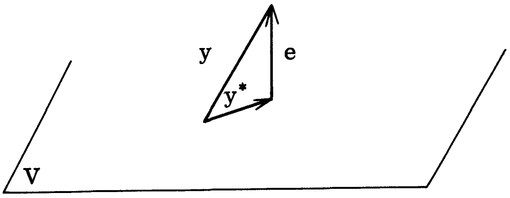
$$y'\underline{y} = y^{*}{'}\underline{y} + e'\underline{e}$$

soit

$$\boxed{\sum_i y_i^2 = \sum_i y_i^{*2} + \sum_i e_i^2}$$

avec

$$\sum_i e_i^2 = \sum_i (y_i - y_i^*)^2$$



4 - Détermination de b

Quand $\underset{(n,p)}{X}$ est de rang p, l'application f associée à X est injective, V est de dimension p et $\underline{b} = f^{-1}(y^*)$ est défini de façon unique.

Si X est de rang r inférieur à p il y a indétermination non pas sur $\underline{y}^* = X\underline{b}$ mais sur $\underline{b}$ qui appartient à la variété $f^{-1}(\underline{y}^*)$ qu'on sait être de dimension p − r.

C'est ce qui se produit quand parmi les variables indépendantes considérées certaines sont qualitatives.

Soit A une variable qualitative prenant K modalités. On lui associe alors K variables binaires $x^{\alpha e}$, $e = 1, 2 \ldots K$ prenant les valeurs 1 où 0 suivant que A prend ou non la modalité A_K. Les K colonnes de X relatives à l'ensemble des $x^{\alpha e}$ forment alors un ensemble de rang K − 1 puisque la somme des éléments d'une ligne pour ces K colonnes est égale à 1. On lève généralement l'indétermination en imposant aux composantes $\beta\alpha_e$, $e = 1, \ldots K$ de β correspondant aux x   de satisfaire à une contrainte, par exemple à $\sum_{e=1}^{K} \alpha_e = 0$ ce qui revient à prendre pour origine des effets des modalités A_e, la moyenne de ces effets.

Equations normales et expression de b

La dérivation par rapport à $\underline{b}$ de $(y - Xb)'M(y - Xb)$ et l'annulation de cette dérivée conduit à $X'M(y - y^*) = 0$ soit $X'MX\underline{b} = X'M\underline{y}$

$X'M(y - y^*) = 0$ traduit la M orthogonalité de $y - y^* = e$ et des colonnes de X donc de $y - y^*$ et de V.

Expression de $\underline{b}$

Dans la suite on supposera que $M = I$

(Sinon on pourra toujours écrire $M = A'A$ et il suffira de poser $\underline{z} = \underline{Ay}$ $U = AX$ pour être ramené au modèle $\underline{z} = U\underline{\beta} + \underline{\eta}$ ($\underline{\eta} = A\underline{\varepsilon}$) que l'on pourra traiter ave la métrique unité. En effet $z'z = y'A'Ay = y'My$)

Si X est de rang p il en est de même de $X'X = S$ et on obtient :

$$\boxed{b = (X'X)^{-1} X'y}$$

Si X est de rang $r < p$, on ajoute les contraintes $\underset{(t,p)}{F} \underline{b} = 0$ avec F de $p - r$ et telle que rang de $G = \left|\left| \begin{matrix} X \\ F \end{matrix} \right|\right| = p$.

$$\text{On a alors} \left\{ \begin{array}{l} X'X\underline{b} = X'\underline{y} \\ \\ F'F\underline{b} = 0 \end{array} \right.$$

soit $G'G\underline{b} = X'\underline{y}$ où $G'G$ est de rang p.

On en déduit

$$\boxed{b = (G'G)^{-1} X'\underline{y}}$$

$\underline{b}$ est un estimateur de $\underline{\beta} = (G'G) \ X'E(\underline{y})$

$$\text{qui satisfait à} \left\{ \begin{array}{l} E(\underline{y}) = X\underline{\beta} \\ F\underline{\beta} \ = 0 \end{array} \right.$$

5 - Propriétés de $\underline{b}$ et des écarts résiduels

a) Théorème de Markov

1/ $\underline{\beta}$ étant défini par $\left\{ \begin{array}{l} E(\underline{y}) = X\underline{\beta} \\ F\underline{\beta} \ = 0 \end{array} \right.$ où F est de rang $p - r$

et telle que $G = \left|\left| \begin{matrix} X \\ F \end{matrix} \right|\right|$ soit de rang p.

$\underline{b}$ est un estimateur sans biais de $\underline{\beta}$ ($E(b) = \beta$) tel que $F\underline{b} = 0$

2/ $\underline{b}$ est fonction linéaire de $\underline{y}$

3/ Si $\text{cov } \varepsilon = \sigma^2 I_n$, parmi les estimateurs $\hat{\beta}$ de $\underline{\beta}$ satisfaisant à 1. et 2., b est tel que ses composantes ont des variances minimales. Plus généralement on a $\text{cov } \underline{\hat{\beta}} - \text{cov } \overline{b}$ définie positive si $\underline{\hat{\beta}} \neq b$. On montre que si X est de rang p : $\text{cov } \underline{\overline{b}} = \sigma^2 (\overline{X'}X)^{-1}$ et dans le cas général :
$\text{cov } \underline{b} = \sigma^2 \ (G'G)^{-1} X'X (G'G)$, avec $G'G = X'X + F'F$

b) On montre aussi que $s^{*2} = \frac{e^2}{n-r} = \frac{\bar{y}^2 - \overline{Xb}^2}{n-r}$ est un estimateur sans biais de σ^2 et que $\underline{e}$ et $\underline{b}$ sont non corrélés ($\underline{e}$ et $X\underline{b}$ sont les projections orthogonales de $\underline{y}$ sur $V\perp$ et V).

c) Propriétés complémentaires quand on suppose que la distribution de ε donc de y est normale de matrice variance $\sigma^2 I_n$

1/ $\underline{b}$ a une distribution normale de rang r caractérisée par sa moyenne $\underline{\beta}$ et sa matrice variance cov $\underline{b}$ (cf a) 3.).

2/ $\underline{e}$ a une distribution normale de rang $n - r$, de moyenne nulle.

3/ e et b sont indépendants en probabilité.

4/ $\underline{e}^2 \in \sigma^2 \chi^2 (n - r)$ et $\| X(\underline{b} - \underline{\beta})\|^2 = \|y^* - E(\underline{y})\|^2 \in \sigma^2 \chi^2_r$,

ces deux variates étant indépendantes en probabilité.

On peut aussi dire que $\overline{Xb}^2 = y^{*2} \in \sigma^2 \chi'^2_r (\overline{X\beta}^2)$ où $\chi'^2_r(\lambda)$ désigne une variable chi deux non centrée à r degrés de liberté de paramètre de non centralité λ.

6 - Test d'une sous-hypothèse linéaire sur les coefficients de régression

Ce qui suit n'est valable que sous l'hypothèse où $\underline{\varepsilon}$ donc $\underline{y}$ suit une distribution normale de matrice variance $\sigma^2 I_n$.

On se propose de tester l'existence de certaines relations linéaires entre les composantes de $\underline{\beta}$.

On peut toujours mettre ces relations sous la forme générale :

$$H(\underline{\beta} - \underline{\theta}) = 0 \tag{4}$$

où H est une matrice (q , p) donnée de rang q et θ est donné.

q est l'ordre de l'hypothèse qu'on désire tester.

Comme exemples les plus fréquemment rencontrés on peut citer l'hypothèse $\underline{\beta}_2 = 0$ où $\underline{\beta}_2$ est une composante de $\underline{\beta} = \left\| \begin{array}{c} \underline{\beta_1} \\ \hline \underline{\beta_2} \end{array} \right\|$ dans la décomposition de $\underline{\beta}$. suivant deux sous-espaces de R^p ayant respectivement $p - q$ et q dimensions. Ce test est alors appelé "test de signification" d'un ensemble de q coefficients de régression partielle.

Plus généralement on peut tester l'hypothèse $\underline{\beta}_2 = \underline{\beta}_{20}$ où $\underline{\beta}_{20}$ est fixé.

On rencontre également assez souvent le test de l'hypothèse d'égalité entre certaines composantes β_j, $\beta_j \in R^s$, de $\underline{\beta}$. On en verra un exemple très important en II.

L'hypothèse (4) signifie que β n'est pas quelconque dans R^p mais qu'il est situé dans une variété linéaire B^{p-q} dimension $p - q$ de R^p.

Ce qu'on teste en fait c'est que $E^x(y) = X\underline{\beta}$ appartient à la variété V^{p-q} image par X de B^{p-q} qui est bien entendu contenue dans l'image V de R^p.

On est donc amené à tester l'hypothèse :

H_1 : (1) $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$

avec (4) $H(\underline{\beta} - \underline{\theta}) = 0$

comme "sous-hypothèse" de H_o : (1) $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$

La méthode générale de test repose sur la comparaison des sommes de carrés résiduelles,

$\bar{e}^2 = Q_{E/H_0}$ et $\bar{e}_1^2 = Q_{E/H_1}$ attachées à l'estimation de $\underline{\beta}$ par la méthode des moindres carrés sous les hypothèses H_0 et H_1 respectivement.

Quand $\theta = 0$, V_1 est un sous-espace de R^n et l'interprétation géométrique de la méthode montre que si $\underline{b}_{H_1}$ est l'estimateur des moindres carrés de $\underline{\beta}$ sous H_1, $\underline{y}_{H_1} = X\underline{b}_{H_1}$ = proj $\underline{y}$ sur V_1

et $\underline{e}_1$ = proj $\underline{y}$ sur $V_1 \perp$

Comme $V_1 \subset V$, $V_1\perp \supset V\perp$. On a donc e = proj e_1 sur $V\perp$.(1)

C'est à partir de cette remarque qu'on montre que la statistique

$$G_{H_1/H_0} = \frac{Q_{E/H_1} - Q_{E/H_0}}{q} : \frac{Q_{E/H_0}}{n - p}$$

est distribuée comme F à q et n − p degrés de liberté si, et seulement si l'hypothèse H_1 est réalisée.

On peut montrer que le test le plus puissant basé sur G_{H_1/H_0} est un test unilatéral à droite.

Si α est le seuil que l'on s'est fixé, on rejettera donc H_1 si

$G_{H_1/H_0} > \underset{q,n-p}{F}(1 - \alpha)$, où $F(1 - \alpha)$ est défini par $P[F < F(1 - \alpha)] = 1 - \alpha$

Remarque : Test de signification d'un coefficient de régression partielle β_j

Dans le test de $H_1 : \beta_j = 0$, $q = 1$ et G_{H_1/H_0} est sous H_1, distribuée comme $F_{1,n-p}$.

On montre que $G_{H_1/H_0} = (b_j/s^*b_j)^2$ où b_j et $s^*_{b_j}$ sont respectivement l'estimateur des moindres carrés de β_j et l'écart type estimé de b_j (obtenu en remplaçant σ inconnu dans σ_{b_j} par s^*) sous l'hypothèse H_0.

L'application de la méthode générale est donc équivalente à l'application du test t de Student à n − p degrés de liberté à $b_j/s^*_{b_j}$.

II - COMPARAISON DES REGRESSIONS OBTENUES SUR K ECHANTILLONS (ANALYSE DE LA COVARIANCE)

On suppose que les formules de régression comportent un terme constant c'est-à-dire que la première colonne de X est formée de 1.

---

(1) Quand $\theta \neq 0$ on se ramène au cas précédent (1) sous la forme
$y - X\theta = X(\beta - \theta) + \varepsilon$
soit (1') $z = X\gamma + \varepsilon$ avec $z = y - X\theta$ connu $\in R^n$
et $\gamma = \beta - \theta \in R^p$
(4) devient alors (4'): $H\gamma = 0$
et on est amené à tester (4') dans le modèle (1')
On aura alors $b_{H_1} = g_{H_1} + \theta$, si g_{H_1} est l'estimateur des moindres carrés de γ.

On est alors amené à se poser les problèmes suivants : test du parallélisme des hypersurfaces de régression : H_1, ou test de la confusion de ces hypersurfaces : H_2

1/ de prime abord, ou

2/ en les supposant parallèles (c'est ce problème qui constitue "l'analyse de la covariance").

Le modèle de départ correspondant à l'hypothèse H_0 est :

$$H_0 \left\{ \begin{array}{ll} \underline{y}_1 = X_1 \underline{\beta}_1 + \underline{\varepsilon}_1 & \underline{\beta}_1,\ \underline{\beta}_2 \ldots \underline{\beta}_K \in R^p \\ \underline{y}_2 = X_2 \underline{\beta}_2 + \underline{\varepsilon}_2 & \underline{y}_i\ ,\ \underline{\varepsilon}_i \in R^{n_i} \\ \text{-------------} & \\ \underline{y}_K = X_K\ \underline{\beta}_K + \underline{\varepsilon}_K & i = 1,\ 2 \ldots K \end{array} \right.$$

Si on fait l'hypothèse que $\varepsilon_i \in N(0\,,\sigma^2 I_{n_i})$ et que les K échantillons sont indépendants, ces problèmes rentrent dans le cadre de la méthode générale de test d'une sous-hypothèse linéaire sur les coefficients de régression (cf. I - 6), mais il ne semble pas qu'ils aient fait l'objet d'un programme sauf en ce qui concerne le test de confusion des hypersurfaces supposées parallèles (H_2 comme sous-hypothèse de H_1).

En effet on peut représenter l'ensemble des observations par le modèle :

$$M : \underline{y} = X\underline{\beta} + \underline{\varepsilon} \qquad\qquad \underline{\varepsilon} \in N(0\,,\sigma^2 I_N)$$

avec

$$y = \left| \begin{array}{c} \underline{y}_1 \\ \\ \underline{y}_K \end{array} \right. \qquad \underline{\varepsilon} = \left| \begin{array}{c} \underline{\varepsilon}_1 \\ \\ \underline{\varepsilon}_K \end{array} \right. \qquad \left\{ \begin{array}{l} y \in R^N \\ \varepsilon \in R^N \\ N = \sum\limits_i n_i \end{array} \right. \qquad \underline{\beta} = \left| \begin{array}{c} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \underline{\beta}_K \end{array} \right. \qquad \beta \in \mathbb{R}^{pk}$$

$$\underset{(N,\, pK)}{X} = \left\| \begin{array}{ccc} X_1 & \ldots\ldots & 0 \\ & X_2 & \\ 0 & & X_k \end{array} \right\|$$

et le test de confusion des hypersurfaces de régression est le test de l'hypothèse H : $\underline{\beta}_1 = \underline{\beta}_2 \ldots\ldots = \underline{\beta}_k$ tandis que le test du parallélisme est celui de l'hypothèse $H_2 : \beta_{1j} = \beta_{2j} \ldots\ldots = \beta_{Kj}\ \forall j \neq 1$

III - CHOIX DES VARIABLES EXPLICATIVES OU INDEPENDANTES A FAIRE INTERVENIR DANS UNE EQUATION DE REGRESSION

1 - Evaluation du biais introduit dans les estimateurs des q coefficients de régression restant quand on supprime p − q variables indépendantes dans un modèle de régression linéaire

Considérons le modèle de régression

(1) $$E^{x'}(y) = x'_1\ \beta_1 + x'_2\ \beta_2 = \pi$$

où $\underline{\beta}_1$ est de dimension q et $\underline{\beta}_2$ de dimension p − q

et le modèle simplifié (2) $E^{x'}(y) = x'_1 \ \beta_1 = \pi_1$

On trouve que si $\underline{b}_1^o$ est l'estimateur des moindres carrés de $\underline{\beta}_1$ dans le modèle (2) on a $E(\underline{b}_1^o) = \underline{\beta}_1 + (X'_1 X_1)^{-1} X'_1 X_2 \underline{\beta}_2$ où X_1 et X_2 sont respectivement les matrices n, q et n, p − q dont les lignes sont les valeurs de $\underline{x'}_1$ et $\underline{x'}_2$ pour le n échantillon considéré ($X = ||X_1|X_2||$ et $\beta = \left|\begin{matrix}\underline{\beta}_1 \\ \underline{\beta}_2\end{matrix}\right.$)

Le biais commis en estimant $\underline{\beta}_1$ par $\underline{b}_1^o$ est donc

$$\underline{B} = (X'_1 X_1) \quad X'_1 X_2 \underline{\beta}_2$$

Il est nul si et seulement si : $\underline{\beta}_2 = 0$ c'est-à-dire si le modèle (2) convient

$$\text{ou} \quad X'_1 X_2 = 0$$

c'est-à-dire si les colonnes de X_1 sont orthogonales aux colonnes de X_2. (on dit si les variables composant X_1 sont orthogonales aux variables composant X_2)

Ce biais dépend donc du plan de relevé des données par l'intermédiaire de X_1 et X_2

On a

$$X_1 \beta = \underbrace{X_1 (X'_1 X_1)^{-1} X'_1 X_2 \beta_2}_{\text{proj}(X_2 \underline{\beta}_2) \text{ sur } V_1}$$

V_1 = sous espace de R^n engendré par les colonnes de X_1

Ce biais est donc l'image réciproque par f_1 associée à X_1 de la projection orthogonale sur V_1 sous espace engendré par X_1, de $X_2 \underline{\beta}_2$ composante de $E^{x'}(y)$ négligée dans le modèle (2).

On montre aussi que $\underline{b}_1^o = \underline{b}_1 + (X'_1 X_1)^{-1} X'_1 \, X_2 \underline{b}_2$

si b_1 et b_2 sont les estimateurs des moindres carrés de β_1 et β_2 dans le modèle (1).

2 - Conclusions de l'étude

1/ La suppression à tort d'une ou plusieurs variables indépendantes conduit à des estimateurs biaisés pour les coefficients de régression des autres variables si celles-ci ne sont pas orthogonales aux variables supprimées dans le n échantillon considéré (ou non corrélées avec elles s'il s'agit de variables centrées).

2/ L'étude de l'influence sur E(y) d'une variable ou d'un groupe de variables considéré isolément exige que cette variable ou ce groupe de variables soit orthogonal à chacune des autres dans l'échantillon considéré. S'il n'en n'est pas ainsi il se peut que l'influence de ces variables puisse être négligée par rapport aux influences des autres quand on se limite à l'ensemble des valeurs prises par les variables indépendantes dans l'échantillon, mais rien ne permet d'affirmer qu'il en sera ainsi pour d'autres valeurs, ou même d'autres combinaisons de valeurs des variables indépendantes. C'est ce qui interdit la suppression de variables indépendantes non orthogonales aux autres pour la prévision par extrapolation, si on ne veut pas risquer d'introduire un biais dont on ne peut pas connaitre la valeur.

De même, dans la prévision par extrapolation de l'influence de certaines variables indépendantes sur E(y) on ne peut envisager que la suppression éventuelle d'un groupe de variables qui seraient orthogonales aux autres dans l'échantillon si on veut prévoir sans biais l'influence des variables restant.

Aussi semblerait-il intéressant de procéder à une analyse en composantes principales sur l'ensemble des variables indépendantes (ou explicatives) pour rechercher des groupes de variables non corrélées (ou plus généralement orthogonales d'un groupe à l'autre si on ne considère pas des variables centrées), ou substituer aux variables initiales de nouvelles variables deux à deux non corrélées ou plus généralement orthogonales. C'est le problème de la "régression orthogonalisée" que nous n'aborderons pas ici.

3 - Les principales procédures de choix des variables indépendantes à faire intervenir dans une équation de régression utilisée pour la prévision par interpolation

On supposera pour simplifier le langage que la formule de régression recherchée comporte un terme constant, ce qui revient à considérer une première variable indépendante x^0 constante, et $p - 1$ variables indépendantes non constantes au sein desquelles on se propose de choisir les variables à conserver. Il importe de ne pas perdre de vue qu'il n'y a pas de procédure statistique unique pour procéder à ce choix et que le jugement personnel devra intervenir chaque fois.

a) Essai de toutes les régressions possibles

Ceci conduit à essayer 2^{p-1} régressions et n'est donc praticable que pour $p - 1 < 10$;

On est amené à choisir "une meilleure régression" dans chaque groupe de régressions à r variables indépendantes et ceci pour toutes les valeurs possibles de r soit $r = 1, 2 \ldots p - 1$ et à comparer ensuite les meilleures régressions obtenues pour les diverses valeurs de r afin de décider quelle est la plus petite valeur de r que l'on retiendra.

Pour procéder à ce choix on utilise deux types de critères :

1/ Le carré du coefficient de corrélation multiple R^2, ou la variance résiduelle empirique $s^2 = e^2/n$

R^2 est la réduction relative de variance introduite par le remplacement de la variance marginale empirique de y par sa variance résiduelle empirique dans la régression considérée.

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - y_i^*)^2}{\sum_i (y_i - y)^2}$$

puisque

$$\bar{e}^2 = \sum_i (y_i - y_i^*)^2$$

On compare donc la variance résiduelle de y attachée à une régression à sa variance en l'absence de toute régression. Comme la variance marginale ne dépend que de l'échantillon, le critère de R^2 maximum équivaut au critère de variance résiduelle empirique minimum.

On notera que quand r augmente par adjonction de nouvelles variables explicatives e^2 ne peut que diminuer donc R^2 ne peut qu'augmenter ; donc la meilleure régression à $r + 1$ termes sera toujours au moins aussi bonne que la meilleure régression à r termes.

Pour choisir la valeur de r à adopter on peut porter les valeurs de R^2 Max ou de e^2 Min/n en fonction de r. On retiendra la plus petite valeur de r pour laquelle R^2Max ou e^2Min/n se stabilise.

2/ Le critère C_K de Mallows

K est le nombre total des termes introduits dans la régression, donc $K = r + 1$ s'il y a r variables explicatives non constantes. Par définition $C_K = K +$ somme des carrés des erreurs systématiques de prévision commises en chacun des n points de l'échantillon, ces erreurs étant comptées en écarts types liés.

K est la valeur moyenne de la somme des carrés des erreurs aléatoires de prévision aux mêmes points, comptée elle aussi en écarts types liés.

On montre que $C = \bar{e}_K^2/\sigma^2 - (n - 2K)$ en désignant par $\bar{e}_K^2$ la somme des carrés résiduelle attachée à la régression à K termes considérée.

Pratiquement on ne connait pas σ^2 et on l'estime par $s^{*2} = \frac{\bar{e}_p^2}{n - p}$ estimateur de σ^2 dans le modèle complet de régression à p termes, en sorte qu'on obtient $C_p = p$

On compare donc en fait la somme des carrés résiduelle des modèles à K termes à la somme des carrés résiduelle du modèle "complet" à p termes.

En l'absence d'erreur systématique $C_K = K$; par suite les valeurs de C_K doivent être comparées à K et on retiendra le plus petit nombre K tel que C_K Min soit voisin de K.

Remarques importantes

- Le meilleur groupe de $r + 1$ variables ne contient pas nécessairement le meilleur groupe de r varaibles, en sorte que même sous l'hypothèse de normalité de la distribution de $\underline{\varepsilon}$ il n'est pas toujours possible de tester la signification de l'influence de la $(r + 1)^{\text{ème}}$ "meilleure" variable.

- Par contre toujours sous l'hypothèse de normalité pour la distribution de $\underline{\varepsilon}$ on peut tester la signification de l'ensemble des $p - (r + 1)$ variables non introduites c'est-à-dire tester l'hypothèse $\beta_2 = 0$ dans le modèle (1) de (I,6) où x' est l'ensemble des $p - (r + 1)$ variables non introduites. Quand on a calculé les coefficients C_K, la méthode générale indiquée en (I,6) conduit à appliquer un test F à $p - K$ et $n - p$ degrés de liberté, unilatéral à $1 + \frac{C_K - K}{p - K}$

b) Elimination progressive (Backward Elimination)

Cette méthode consiste à :

1/ Calculer une formule de régression comportant toutes les variables.

2/ Calculer la valeur de toutes les statistiques

$$\frac{b_j}{s^*_{b_j}} \quad , \; j = 1,2 \ldots p - 1$$

qui permettent de tester la signification des régressions partielles β_j au moyen du test t de Student (cf I, 6 : Remarque)

3/ Tester la signification de la régression partielle la moins significative (celle qui correspond à la plus petite valeur de $b_j / s^*_{b_j}$) par exemple de β_L .

4/ - Si cette régression partielle est significative, conserver la formule trouvée.

- Si elle n'est pas significative, supprimer la variable x_L et recommencer le processus.

Cette méthode est particulièrement recommandée à ceux qui désirent avoir la formule de régression sur l'ensemble des variables indépendantes. Elle donne par ailleurs la meilleure régression à $p - 1$ termes.

Si on ne tient pas spécialement à garder ou du moins à essayer l'ensemble des $p - 1$ variables indépendantes, la régression pas à pas parait préférable.

Remarques

1/ Le critère de choix de la variable à éliminer à chaque étape est en fait le carré du coefficient de corrélation partielle de y avec cette variable.

Il garde un sens concret même si les hypothèses de validité du test appliqué ne sont pas réalisées.

2/ Il faut se garder d'éliminer simultanément plusieurs variables qui auraient des régressions partielles non significatives ; c'est en fait souvent parce que ces variables sont fortement corrélées qu'aucune d'elles n'a d'influence significative.

3/ On peut envisager, une fois la formule de régression finale obtenue, de tester la signification de l'ensemble des variables éliminées.

c) Régression pas à pas ou "Stepwise regression"

(cf B.M.D. 02 - R)

C'est une méthode d'adjonction progressive des variables.

Les étapes de la méthode sont les suivantes :

1/ Choisir la variable x_j la plus fortement corrélés avec y, soit x_1 cette variable.

Tester la signification du coefficient de régression totale b_1 à un seuil α_1 fixé.

S'il n'est pas significatif, conclure à l'absence de régression significative de y sur les x_j.

Si l'influence de x_1 est significative passer à

2/ Introduire successivement chacune des variables x_j, $j \neq 1$, en plus de x_1.

Retenir la variable qui donne un coefficient de corrélation multiple $R^2 y/x_1,x_j$ le plus grand possible c'est-à-dire une somme des carrés résiduelle $\bar{e}^2$ la plus petite possible.

Soit x_2 cette variable. Tester la signification du coefficient de régression partielle β_2 dans la régression de y sur x_1 et x_2 au risque α_1.

Si l'influence de x_2, après x_1 n'est pas significative, garder la seule variable explicative x_1.

Si elle est significative passer à

3/ Tester la signification de x_1, après x_2 au risque α_2 $(\alpha_2 \geqslant \alpha_1)$.

a) Si x_1 n'est pas significatif après x_2, garder x_2 seule et reprendre à l'étape 2. Le fait que $\alpha_2 \geqslant \alpha_1$ implique que x_1 ne peut avoir après x_2 une influence significative au risque α_1 et ne peut par suite pas être réintroduite après x_2

b) Si x_1 est significatif après x_2, garder x_1 et x_2 et passer à

4/ Introduire successivement chacune des variables x_j, $j \geqslant 3$ en plus de x_1 et x_2 et retenir la variable donnant la plus forte corrélation multiple.

Soit x_3 cette variable. Si son influence après x_1 et x_2 n'est pas significative au risque α_1, garder la formule avec x_1 et x_2. Si l'influence de x_3 après x_1 et x_2 est significative, tester au risque α_2 la signification de celle des variables x_1 et x_2 qui est la moins significative ($b_j/s^*_{b_j}$ le plus petit).

a) Si les deux variables x_1 et x_2 ont des influences significatives, poursuivre le processus en essayant d'introduire une quatrième variable.

b) Si x_1 par exemple n'est pas significative, la supprimer et passer à l'étape 4.

On arrête le processus quand il n'y a plus aucune variable à introduire ou à supprimer.

La méthode est aisée à mettre en oeuvre si on fait apparaitre les $s^*_{b_j}$ à chaque étape, c'est-à-dire si on calcule les éléments diagonaux de $(X'X)^{-1}$ puisque le test de signification d'une régression partielle β_j est le test t appliqué à b_j/s_{b_j}. Le plus souvent on ne change pas de valeur critique pour t quand le nombre K des variables augmente. C'est sans importance quand n est grand devant K.

Remarques

1/ Cette méthode conduit à renoncer à toute régression dans le cas où un ensemble de variables serait collectivement significatif sans qu'aucune d'elle considérée isolément ne le soit.

On peut se prémunir contre ce risque en calculant la régression sur l'ensemble des p − 1 variables condidates et en testant la signification de la régression globale ou de façon plus concrète en calculant le coefficient de corrèlation multiple ou le coefficient C_1 de Mallows correspondant au seul terme constant.

D'une façon générale, une fois l'application de la méthode achevée, on peut encore tester la signification de l'ensemble des variables éliminées ou calculer le coefficient C_K correspondant à la formule trouvée.

2/ Quand on a plus de variables candidates que d'observations, cette méthode fournit un bon moyen de choix d'un nombre raisonnable de variables.

Il peut être intéressant d'appliquer simultanément les deux dernières méthodes et d'en comparer les résultats. Quand on obtient la même régression avec les deux méthodes, on peut penser que celle-ci est satisfaisante. Dans le cas contraire, c'est que le choix d'un sous-ensemble de l'ensemble des variables candidates présente des difficultés et il convient de faire davantage appel à la réflexion.

REFERENCES BIBLIOGRAPHIQUES POUR LA PARTIE III

DRAPER & SMITH - Applied Regression Analysis, Wiley 1966 - chapitre VI

H.C. HAMAKER - De l'analyse à régression multiple - R.S.A. - 1962, vol.1 - p. 23 - 48

Pour le critère C_K de Mallows

GORMAN & TOMAN - Selection of Variables For Fitting Equations to Data, Technometrics - Vol. 8 - N° 1 - 1966 - p. 27-51.

SUR LA REGRESSION LINEAIRE EN GENERAL

J. ULMO - Etude fondamentale de la régression linéaire multiple - Ronéoté - Tome I. (I.S.U.P.)

F.A. GRAYBILL - An introduction to linear statistical Models. Volume I. Mac Graw Hill. 1961