SY09
Régression linéaire

T. Denœux

1 Principe

Nous avons vu dans le chapitre sur la théorie de la décision que la fonction de décision g minimisant le risque quadratique

$$R(g) = \mathbb{E}_{\mathbf{X},Y}[(g(\mathbf{X}) - Y)^2]$$

est la fonction de régression :

$$g^*(\mathbf{x}) = \mathbb{E}(Y|\mathbf{x}).$$

Dans le modèle de régression linéaire, on suppose que cette fonction est une fonction affine de $\mathbf{x}$ que l'on note

$$g^*(\mathbf{x}) = w_0^* + \sum_{i=1}^{p} w_j^* x_j = \mathbf{w}^{*\prime}\mathbf{x}$$

avec, par convention, $\mathbf{x} = (1, x_1, \ldots, x_p)'$.

Le problème posé ici consiste à estimer le vecteur $\mathbf{w}^*$ à partir d'un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Pour cela, on remarque que $\mathbf{w}^*$ s'obtient comme solution d'un problème d'optimisation :

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} R(\mathbf{w})$$

avec $R(\mathbf{w}) = \mathbb{E}_{\mathbf{X},Y}[\mathbf{w}'\mathbf{X} - Y)^2]$. On ne peut en pratique résoudre ce problème de manière exacte car la fonction $R(\mathbf{w})$ est inconnue, mais on peut remplacer le risque théorique par le *risque empirique* défini par :

$$\widehat{R}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}'\mathbf{x}_i - y_i)^2.$$

On notera $\widehat{\mathbf{w}}$ le vecteur de coefficients minimisant le risque empirique :

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \widehat{R}(\mathbf{w}).$$

Intuitivement, lorsque n est assez grand, $\widehat{R}(\mathbf{w})$ sera "proche" de $R(\mathbf{w})$, et $\widehat{\mathbf{w}}$ sera donc "proche" de $\mathbf{w}^*$. Le vecteur $\widehat{\mathbf{w}}$ est appelé *estimateur des moindres carrés* de $\mathbf{w}^*$.

2 Méthode des moindres carrés

Notons

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

la matrice $(n, p+1)$ contenant les valeurs des variables explicatives et $\mathbf{y} = (y_1, \dots, y_n)'$ le vecteur des observations de la variable y. Le risque empirique $\widehat{R}(\mathbf{w})$ peut alors s'écrire matriciellement :

$$\begin{aligned} \widehat{R}(\mathbf{w}) &= \frac{1}{n}(X\mathbf{w} - \mathbf{y})'(X\mathbf{w} - \mathbf{y}) \\ &= \frac{1}{n}\left(\mathbf{w}'X'X\mathbf{w} - 2\mathbf{w}'X'\mathbf{y} + \mathbf{y}'\mathbf{y}\right). \end{aligned}$$

On a

$$\frac{d\widehat{R}(\mathbf{w})}{d\mathbf{w}} = \frac{1}{n}\left(2X'X\mathbf{w} - 2X'\mathbf{y}\right),$$

d'où

$$\frac{d\widehat{R}(\mathbf{w})}{d\mathbf{w}} = 0 \Leftrightarrow X'X\mathbf{w} = X'\mathbf{y}. \tag{1}$$

Si la matrice $X'X$ est inversible, le minimum du risque empirique est donc obtenu pour

$$\widehat{\mathbf{w}} = (X'X)^{-1}X'\mathbf{y}.$$

On notera

$$\widehat{\mathbf{y}} = X\widehat{\mathbf{w}} = X(X'X)^{-1}X'\mathbf{y}$$

le vecteur des prédictions obtenu en remplaçant le paramètre $\mathbf{w}$ inconnu par son estimateur des moindres carrés $\widehat{\mathbf{w}}$.

Remarque 1 *On a supposé $X'X$ inversible, ce qui est le cas si la matrice X est de rang $p+1$. Si ce n'est pas le cas, c'est qu'une variable (une colonne de X) s'exprime comme combinaison linéaire des autres. Il suffit alors de supprimer la ou les variables redondantes.*

Remarque 2 *Si certaines variables sont très corrélées, la matrice $X'X$ est mal conditionnée est les calculs numériques peuvent être très imprécis. Une solution (appelée* ridge regression *en anglais) consiste à ajouter un terme sur la diagonale de $X'X$:*

$$\widehat{\mathbf{w}}_\lambda = (X'X + \lambda I)^{-1}X'\mathbf{y}$$

où λ est une constante à déterminer. On montre que l'on améliore ainsi parfois les propriétés de l'estimateur.

Remarque 3 *On a*

$$\widehat{\mathbf{y}} = X\widehat{\mathbf{w}} = X(X'X)^{-1}X'\mathbf{y} = P\mathbf{y}$$

en notant $P = X(X'X)^{-1}X'$. Cette matrice P a des propriétés remarquables. En effet, P est symétrique (évident), et de plus

$$P^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P.$$

La matrice P est donc idempotente (c'est un opérateur de projection orthogonale, comme nous le verrons par la suite). De même, on peut écrire

$$\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}} = (I_n - P)\mathbf{y} = R\mathbf{y}$$

avec $R = I_n - P$. On vérifie aisément que R a les mêmes propriétés que P (symétrie et idempotence) : c'est également un opérateur de projection orthogonale.

3 Analyse de la variance

3.1 Point de vue géométrique

Plaçons nous dans $\mathbb{R}^n$ et considérons les vecteurs

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \ j = 1, p$$

La méthode des moindres carrés peut être interprétée comme la recherche de la meilleure approximation de $\mathbf{y}$ dans le sous-espace $\mathcal{L}$ de $\mathbb{R}^n$ engendré par les $p+1$ vecteurs $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$. On cherche en effet

$$\widehat{\mathbf{y}} = \widehat{w}_0 \mathbf{1} + \sum_{j=1}^{p} \widehat{w}_j \mathbf{x}_j \in \mathcal{L}$$

tel que la distance euclidienne $\|\mathbf{y} - \widehat{\mathbf{y}}\|^2$ soit minimum. On sait que la solution consiste à définir $\widehat{\mathbf{y}}$ comme la projection orthogonale de $\mathbf{y}$ sur $\mathcal{L}$. On a vu en effet que

$$\widehat{\mathbf{y}} = P\mathbf{y},$$

P étant un opérateur de projection orthogonale.

Cette représentation géométrique permet de retrouver sans calculs fastidieux plusieurs résultats intéressants. Tout d'abord,

$$\widehat{\varepsilon} \perp \mathbf{1} \Rightarrow \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0,$$

d'où l'on déduit

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}.$$

Par ailleurs, la projection orthogonale de $\mathbf{y}$ sur l'axe dirigé par $\mathbf{1}$ a pour coordonnée

$$\frac{<\mathbf{y}, \mathbf{1}>}{\|\mathbf{1}\|} = \overline{y}.$$

Il en est de même, d'après ce qui précède, pour la projection orthogonale de $\widehat{\mathbf{y}}$ sur $\mathbf{1}$. Enfin, on a de manière évidente :

$$\widehat{\mathbf{y}} \perp \widehat{\varepsilon}.$$

3.2 Equation d'analyse de la variance

Notons $\overline{\mathbf{y}} = \overline{y}\mathbf{1}$. En appliquant le théorème de Pythagore au triangle $(\mathbf{y}, \widehat{\mathbf{y}}, \overline{\mathbf{y}})$, on obtient finalement la relation très importante suivante, appelée *équation d'analyse de la variance* :

$$\|\mathbf{y} - \overline{\mathbf{y}}\|^2 = \|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2 + \|\widehat{\varepsilon}\|^2,$$

ce que l'on peut encore écrire, en divisant par n :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

soit encore

$$S_{YY} = S_{reg} + S_{res}.$$

Cette équation est appelée *équation d'analyse de la variance.* Le terme de gauche (S_{YY}) est la variance empirique des Y_i, il caractérise la dispersion des valeurs observées de la variable à expliquer. Le premier terme du membre de droite (S_{reg}) est la variance empirique des $\widehat{Y}_i$, que l'on appelle variance expliquée par la régression. Le second terme du membre de droite (S_{res}) est la variance des résidus, ou variance résiduelle.

Remarque 4 *A chacun des termes de l'équation d'analyse de la variance est associé un nombre de degrés de liberté (d.d.l.), égal au nombre de combinaisons linéaires des Y_i utilisées dans le calcul :*

- *S_{YY} dépend de n quantités $y_1 - \overline{y}, \ldots, y_n - \overline{y}$ liées par la relation*

$$\sum_{i=1}^{n}(y_i - \overline{y}) = 0.$$

  Ce terme a donc $n-1$ d.d.l.
- *On a $\widehat{y}_i = \mathbf{x}_i'\widehat{\mathbf{w}}$ et $\overline{y} = \overline{\mathbf{x}}'\widehat{\mathbf{w}}$. Par conséquent, le terme S_{reg} est fonction des paramètres $\widehat{w}_1, \ldots, \widehat{w}_p$ (le terme $\widehat{w}_0$ s'annule dans chacune des différences $\widehat{y}_i - \overline{y}$). La variance expliquée a donc p d.d.l.*
- *Par conséquent, le nombre de d.d.l associé à la variance résiduelle est $n-p-1$.*

La plupart des logiciels statistiques présentent les résultats de la régression sous forme d'un tableau (appelé *tableau d'analyse de la variance*), où figurent les différents termes de l'équation d'analyse de la variance, et les nombres de d.d.l associés (cf. tableau 1).

3.3 Evaluation de la qualité de l'ajustement

On définit à partir de l'équation d'analyse de la variance le *coefficient de détermination*, égal à la proportion de la variance totale expliquée par la régression :

$$R^2 = \frac{S_{reg}}{S_{YY}} = 1 - \frac{S_{res}}{S_{YY}}.$$

Ce coefficient traduit la « qualité de l'ajustement », comme on le voit en considérant les deux situations extrêmes suivantes :

TABLE 1 – Tableau d'analyse de la variance (SS : *sum of squares*; MS : *mean square*).

source de variation	d.d.l.	SS	MS=SS/d.d.l.
régression	p	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$	$\frac{1}{p}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$
résiduelle	$n-p-1$	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \widehat{\sigma}^2$
totale	$n-1$	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

- Si les résidus sont nuls, on a $S_{res} = 0$ et $R^2 = 1$. Les n points $(\mathbf{x}_i, y_i) \in \mathbb{R}^{p+1}$ sont alors situés dans l'hyperplan d'équation

$$y = \widehat{w}_0 + \widehat{w}_1 x_1 + \ldots + \widehat{w}_p x_p.$$

  Cela signifie que l'on peut retrouver sans erreur les y_i à partir des $\mathbf{x}_i$, c'est-à-dire que toute la variation des y_i est expliquée par les $\mathbf{x}_i$.
- Si les prédictions sont constantes ($\widehat{y}_i = \overline{y}, \forall i$), la variance expliquée est nulle et $R^2 = 0$. Dans ce cas, les $\mathbf{x}_i$ n'expliquent pas du tout la variation des y_i.
- De manière générale, on a $0 \leq R^2 \leq 1$, et la valeur du R^2 s'interprète comme un « degré de liaison » entre les variables explicatives et la variable à expliquer.

Remarque 5 *Géométriquement, R^2 est égal au carré du cosinus de l'angle θ entre les vecteurs $\mathbf{y} - \overline{\mathbf{y}}$ et $\widehat{\mathbf{y}} - \overline{\mathbf{y}}$: c'est donc le carré du coefficient de corrélation linéaire entre les y_i et les $\widehat{y}_i$.*