La régression multiple

Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

March 28, 2011

Plan

Rappel sur la régression linéaire

x hauteur (en m) des arbres dominants

y logarithme du nombre de nids de processionnaires par arbre

le modèle :

$$y = ax + b + \varepsilon$$

on cherche à expliquer la variable y par la variable x

on dispose d'un échantillon (d'une ensemble d'observations)

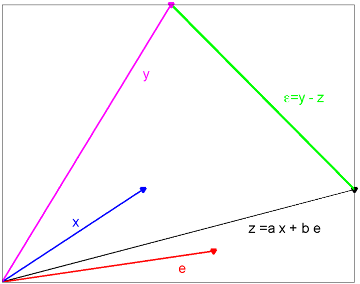
$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n),$$

on cherche à estimer a et b à partir de l'échantillon

Version vectorielle dans $\mathbb{R}^n$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Moyenne :
  - $\mathbf{x}^\top \mathbf{e} = \sum x_i = n\,\bar{x}$
  - $\mathbf{y}^\top \mathbf{e} = \sum y_i = n\,\bar{y}$
- Variance – covariance
  - $||\mathbf{x} - \bar{x}\,\mathbf{e}||^2 = n\,V(\mathbf{x})$
  - $||\mathbf{y} - \bar{y}\,\mathbf{e}||^2 = n\,V(\mathbf{y})$
  - $(\mathbf{x} - \bar{x}\,\mathbf{e})^\top(\mathbf{y} - \bar{y}\,\mathbf{e}) = n\,\mathrm{cov}(\mathbf{x},\mathbf{y})$
- Régression : pour tout couple (a, b)
  - $\mathbf{z} = a\mathbf{x} + b\mathbf{e}$
  - $\varepsilon = \mathbf{y} - \underbrace{a\mathbf{x} + b\mathbf{e}}_{\mathbf{z}}$



l'ensembles des « modèles »

un modèle = un couple (a, b) $\left\{ \begin{array}{ll} y_1 & = ax_1 + b + \varepsilon_1 \\ \vdots & \quad\vdots \\ y_i & = ax_i + b + \varepsilon_i \\ \vdots & \quad\vdots \\ y_n & = ax_n + b + \varepsilon_n \end{array} \right.$ a et b scalaires;

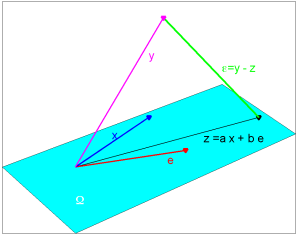
$\mathbf{x}, \mathbf{e}, \varepsilon$, $\mathbf{y}$ et $\mathbf{z}$ vecteurs de $\mathbb{R}^n$

$$\mathbf{y} = \underbrace{a\mathbf{x} + b\mathbf{e}}_{\mathbf{z}} + \varepsilon$$

le plan engendré par les vecteurs $\mathbf{x}$ et $\mathbf{e}$

$$\Omega = \left\{ \mathbf{z} \in \mathbb{R}^n \mid \exists (a, b) \in \mathbb{R}^2 , \ \mathbf{z} = a\mathbf{x} + b\mathbf{e} \right\}$$

$$\mathbf{x}, \mathbf{e} \text{ et } \mathbf{z} \in \Omega$$



Quel est le meilleur modèle?

Présentation matricielle

$$X = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_i & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$$ matrice des observations

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$ la variable à expliquer et

$$\alpha = \begin{bmatrix} a \\ b \end{bmatrix}$$ le vecteur des inconnues

écriture scalaire

$$\begin{cases} ax_1 + b + \varepsilon_1 & = y_1 \\ \vdots & \vdots \\ ax_i + b + \varepsilon_i & = y_i \\ \vdots & \vdots \\ ax_n + b + \varepsilon_n & = y_n \end{cases}$$

a et b sont des scalaires; $\mathbf{x}, \mathbf{e}, \varepsilon$ et $\mathbf{y}$ sont des vecteurs de $\mathbb{R}^n$

$$a\,\mathbf{x} + b\,\mathbf{e} + \varepsilon = \mathbf{y}$$

écriture matricielle

$$X\alpha + \varepsilon = \mathbf{y}$$

Version matricielle du cout

Le problème de moindres carré se réécrit de la manière suivante :

$$\min_{\alpha} J(\alpha) \qquad \text{avec} \quad J(\alpha) = \frac{1}{2}||\mathbf{y} - X\alpha||^2$$

soit en développant :

$$\begin{aligned}
\min_{\alpha} J(\alpha) &= \tfrac{1}{2}||\mathbf{y} - X\alpha||^2 \\
&= \tfrac{1}{2}(\mathbf{y} - X\alpha)^\top(\mathbf{y} - X\alpha) \\
&= \tfrac{1}{2}\mathbf{y}^\top\mathbf{y} - \tfrac{1}{2}\alpha^\top X^\top\mathbf{y} - \tfrac{1}{2}\mathbf{y}^\top X\alpha + \tfrac{1}{2}\alpha^\top X^\top X\alpha \\
&= \tfrac{1}{2}\mathbf{y}^\top\mathbf{y} - \alpha^\top X^\top\mathbf{y} + \tfrac{1}{2}\alpha^\top X^\top X\alpha
\end{aligned}$$

Calcul matriciel : $\mathbf{y}^\top\mathbf{y}$ est un scalaire, $X^\top\mathbf{y}$ est un vecteur de $\mathbb{R}^2$ et $X^\top X$ est une matrice 2×2.

Dérivées partielles du coût

$$
\begin{array}{ccccccccc}
J(\alpha) & = & \frac{1}{2}\mathbf{y}^\top\mathbf{y} & - & \alpha^\top X^\top \mathbf{y} & + & \frac{1}{2}\alpha^\top X^\top X\alpha \\
\frac{\partial J(\alpha)}{\partial \alpha_i} & = & 0 & - & w_i & + & \frac{1}{2}\sum_{j=1}^{p+1}(M_{ij}+M_{ji})\alpha_j
\end{array}
$$

avec $\mathbf{w} = X^\top \mathbf{y}$ et $M = X^\top X$

- $\dfrac{\partial \alpha^\top \mathbf{w}}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{p+1} w_j\alpha_j}{\partial \alpha_i} = w_i$
- $\dfrac{\partial \alpha^\top M\alpha}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{p+1}\sum_{k=1}^{p+1}\alpha_j\alpha_k M_{jk}}{\partial \alpha_i} = \displaystyle\sum_{j=1}^{p+1}\alpha_j M_{ji} + \sum_{k=1}^{p+1}\alpha_k M_{ik}$

car $(uv)' = uv' + u'v$ avec $u = \alpha_j$ et $v = \sum_{k=1}^{p+1}\alpha_k M_{jk}$

$$\nabla J(\widehat{\alpha}) = -\mathbf{w} + M\alpha \qquad = -X^\top \mathbf{y} + X^\top X\widehat{\alpha}$$

La minimisation du cout

La minimisation du cout $J(\alpha)$ est réalisée lorsque le gradient du cout s'annule, soit lorsque :

$$\nabla J(\widehat{\alpha}) = 0 \Leftrightarrow -X^\top \mathbf{y} \; + \; X^\top X \widehat{\alpha} = 0$$

C'est-à-dire dans le cas de la régression simple :

$$-\begin{pmatrix} xy_1 \\ xy_2 \end{pmatrix} + \begin{pmatrix} xx_{11} & xx_{12} \\ xx_{21} & xx_{22} \end{pmatrix} \begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

La solution du problème de minimisation des moindres carrés est le vecteur $\widehat{\alpha}$ défini par :

$$\widehat{\alpha} = \left(X^\top X\right)^{-1} X^\top \mathbf{y}$$

Prévision

$$
\begin{aligned}
z &= X \qquad \widehat{\alpha} \\
&= X \left(X^\top X\right)^{-1} X^\top \mathbf{y} \\
&= \qquad H \qquad \mathbf{y}
\end{aligned}
$$

avec $H = X \left(X^\top X\right)^{-1} X^\top$

Theorem (les moindres carrés comme une projection)

H est un projecteur de $\mathbb{R}^n$

démonstration : $HH = H$ puisque

$$
\begin{aligned}
HH &= X \left(X^\top X\right)^{-1} X^\top \quad X \left(X^\top X\right)^{-1} X^\top \\
&= X \left(X^\top X\right)^{-1} \left(X^\top X \left(X^\top X\right)^{-1}\right) X^\top \\
&= X \left(X^\top X\right)^{-1} X^\top \qquad\qquad = H
\end{aligned}
$$

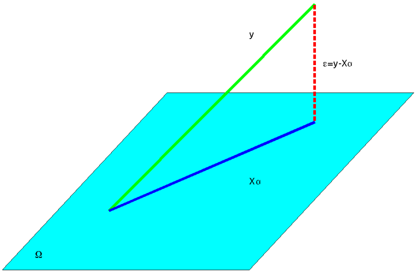
Plan

La régression du point de vue géométrique

- $\Omega = \text{span}\{\mathbf{x}, \mathbf{e}\}$

  le sous espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs $\mathbf{x}$ et $\mathbf{e}$ (c'est-à-dire les colonnes de la matrice X)

  $\mathbf{z} \in \Omega \Leftrightarrow \exists\ a, b \in \mathbb{R} \quad \mathbf{z} = a\mathbf{x} + b\mathbf{e}$



La régression est la projection du vecteur $\mathbf{y}$ sur le sous espace vectoriel engendré par les observations X. En effet, on a :

$$X\widehat{\alpha} = H\mathbf{y} \qquad \text{avec} \qquad H = X\left(X^\top X\right)^{-1} X^\top \tag{1}$$

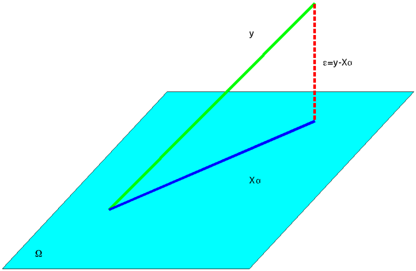
Estimation et orthogonalité

$$J(\alpha) = \frac{1}{2}\,\|\varepsilon\|^2$$

$\|\varepsilon\|^2$ est minimal quand on choisi α tel que $\mathbf{z} = X\alpha$ est la projection orthogonale de $\mathbf{y}$ sur Ω

$$\forall \mathbf{v} \in \Omega \quad \mathbf{v}^\top \varepsilon = 0$$

$$X^\top(\mathbf{y} - X\widehat{\alpha}) = 0$$



$$\begin{aligned} X^\top(\mathbf{y} - X\widehat{\alpha}) = 0 \quad & \Leftrightarrow \quad X^\top\mathbf{y} - X^\top X\widehat{\alpha} = 0 \\ & \Leftrightarrow \quad X^\top X\widehat{\alpha} = X^\top\mathbf{y} \\ & \Leftrightarrow \quad \widehat{\alpha} = (X^\top X)^{-1}X^\top\mathbf{y} \end{aligned}$$

Plan

La Régression multiple

- Variables explicatives :
  1. altitude (en m)
  2. pente (en degré)
  3. nombre de pins moyens dans une placette de 5 ares
  4. hauteur de l'arbre échantillonné au centre de la placette
  5. diamètre de cet arbre
  6. note de densité du peuplement
  7. orientation de la placette (1 : sud, 2 : autres)
  8. hauteur (en m) des arbres dominants
  9. nombre de strates de végétation
  10. mélange du peuplement (1 : pas mélangé, 2 : mélangé)
- Variable à expliquer (y) :
  - logarithme du nombre de nids de processionnaires par arbre d'une placette

source : Tomassone et al. [1992]

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots a_j x_j + \dots + a_p x_p + \varepsilon$$

$p = 1$ c'est la régression linéaire simple

Les objectifs de la régression multiple

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots a_j x_j + \cdots + a_p x_p + \varepsilon$$

- identifier le modèle :
  - estimer les coefficients $a_j,\ j = 0, p$ à partir des observations disponibles
  - estimer la précision de cette estimation
  - décider de la nature de l'influence de chaque variable explicative x_j sur la variable à expliquer y (si en fait $a_j = 0$)
- valider le modèle :
  - donner une mesure de la qualité globale de la régression
- détecter les points aberrants où hors épure
- prévoir pour une nouvelle observation $\mathbf{x}$ la valeur de y

Les données

X_1	X_2	X_3	...	X_10	Y
-0.6416	-0.3934	-0.7313	...	0.9820	0.5524
0.8415	2.1730	0.3442	...	1.6411	0.7336
0.6521	0.5006	3.4307	...	-0.5842	-1.6001
0.1848	-0.2535	0.3812	...	-0.8544	-0.3328
0.3635	0.5542	1.9298	...	0.6564	0.6291
0.5890	-0.2903	1.5685	...	-1.6603	-0.5648
0.5153	-0.4606	0.2263	...	1.6934	-1.0449
0.9598	1.2378	-2.0753	...	0.5395	-0.4543
1.9332	2.2013	-0.2789	...	0.2030	-1.4092
-0.0956	-2.0578	-0.9055	...	-2.3387	1.0096
0.5377	2.7694	1.4090	...	-0.3034	0.3252
1.8339	-1.3499	1.4172	...	0.2939	-0.7549
-2.2588	3.0349	0.6715	...	-0.7873	1.3703
0.8622	0.7254	-1.2075	...	0.8884	-1.7115
0.3188	-0.0631	0.7172	...	-1.1471	-0.1022
-1.3077	0.7147	1.6302	...	-1.0689	-0.2414
-0.4336	-0.2050	0.4889	...	-0.8095	0.3192
0.3426	-0.1241	1.0347	...	-2.9443	0.3129
3.5784	1.4897	0.7269	...	1.4384	-0.8649

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} & \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

$$p = 10, \qquad n = 19$$

attention

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

n lignes et $p + 1$ colones

La Régression multiple : le modèle matriciel

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots a_j x_j + \dots + a_p x_p + \varepsilon$$

$$\mathbf{y} = X\alpha + \varepsilon$$

Hypothèse

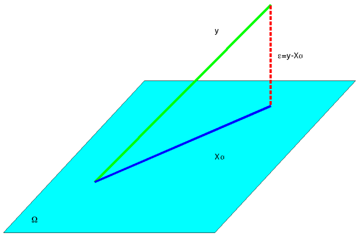
X est une matrice de rang $p+1$ et donc la matrice $X^\top X$ est inversible

$\|\varepsilon\|^2$ est minimal quand on choisi α tel que $\mathbf{z} = X\alpha$ est la projection orthogonale de $\mathbf{y}$ sur Ω

$\forall \mathbf{v} \in \Omega \quad \mathbf{v}^\top \varepsilon = 0 \qquad X^\top(\mathbf{y} - X\widehat{\alpha}) = 0$

$$\boxed{\widehat{\alpha} = (X^\top X)^{-1} X^\top \mathbf{y}}$$

on doit résoudre un système linéaire de $p+1$ inconnues et $p+1$ équations



Mise en œuvre

$$\boxed{\widehat{\alpha} = (X^\top X)^{-1} X^\top \mathbf{y}}$$

Résoudre un système de $p+1$ équations linéaires à $p+1$ inconnues.

- méthode directe
  - calcul de l'inverse : $M = \text{inv}(X^\top X)$
  - calcul des alpha : $\widehat{\alpha} = M(X^\top \mathbf{y})$
- factorisation de Cholesky
  - $\widehat{\alpha} = (X^\top X) \backslash (X^\top \mathbf{y})$
- factorisation QR
  - $X = QR$, Q matrice orthogonale, et R, une matrice triangulaire supérieure
  - $\left(R^\top Q^\top QR\right) \widehat{A} = R^\top Q^\top \mathbf{y}$ soit
  - $R\widehat{\alpha} = Q^\top \mathbf{y}$

Solution Matlab

```
function ahat = regression(x,y)

n = length(y);                  % nombre d'observations
X = [ones(n,1) x];              % construction du regresseur
% ahat = inv(X'*X)*(X'*yi); % calcul direct de l'inverse de la matrice
ahat = (X'*X)\(X'*yi);      % résolution du système linéaire par cholesky

% [Q R] = qr(X);
% ahat = R\(Q'*yi);             % résolution par la méthode de Householder
% ahat = X\yi;                  % idem mais directement
%-----------------------------------------------
function TestRegression

n = 10;= d = 5;
xi = rand(n,d); X = [ones(n,1) xi];
a = randn(d+1,1);
sig = 0.1;
yi = X*a + sig*randn(n,1);

ahat = regression(xi,yi);

erreur = norm(a-ahat);
```

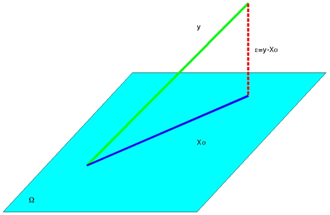
Qualité du modèle : Décomposition de la variance

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-z_i+z_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-z_i)^2+\sum_{i=1}^{n}(z_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-z_i)(z_i-\bar{y}) \\
&= \sum_{i=1}^{n}\varepsilon_i^2+\sum_{i=1}^{n}(z_i-\bar{y})^2+2\sum_{i=1}^{n}\varepsilon_i\Big(\sum_{j=1}^{p+1}x_{ij}a_j-\bar{y}\Big)
\end{aligned}
$$

$$
\boxed{\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SC Total}} = \underbrace{\sum_{i=1}^{n}(y_i-z_i)^2}_{\text{SC Résiduels}} + \underbrace{\sum_{i=1}^{n}(z_i-\bar{y})^2}_{\text{SC Expliqué}}}
$$

Interprétation graphique :

$$
\begin{aligned}
\|y\|^2 &= \|\varepsilon\|^2+\|z\|^2 \\
\|y-\bar{y}\|^2 &= \|\varepsilon\|^2+\|z-\bar{y}\|^2
\end{aligned}
$$



Qualité du modèle : mesurée par le R^2

$$\boxed{\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SC Total}} = \underbrace{\sum_{i=1}^{n}(y_i-z_i)^2}_{\text{SC Résiduels}} + \underbrace{\sum_{i=1}^{n}(z_i-\bar{y})^2}_{\text{SC Expliqué}}}$$

le coefficient de détermination R^2

$$0 \leq R^2 \leq 1$$

$$\begin{aligned} R^2 &= \frac{\text{SC Expliqué}}{\text{SC Total}} = \frac{\sum_{i=1}^{n}(z_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \\ &= \cos^2(\mathbf{y}-\bar{y}\mathbf{e}, \mathbf{z}-\bar{y}\mathbf{e}) \\ &= 1 - \frac{\text{SC Résiduels}}{\text{SC Total}} \end{aligned}$$

R^2 proche de 1 :
le modèle est bon

R^2 proche de 0 :
le modèle est mauvais

le coefficient de corrélation multiple : $R = \sqrt{R^2} = |\mathrm{cor}(\mathbf{y}, \mathbf{z})|$; $\quad \mathbf{z} = X\widehat{\alpha}$

Régression sur variables qualitatives

y : le rendement (quintaux/ha) = rendement moyen + effet variété

$$\begin{aligned} y &= a_0 \quad +a_1x_1 \quad +a_2x_2 \quad +\dots+a_jx_j \quad +\dots \quad +a_px_p \\ &= a_0 \quad +a_1 0 \quad +a_2 0 \quad +\dots+a_j 1 \quad +\dots \quad +a_p 0 \quad = a_0 + a_j \end{aligned}$$

$$X = \begin{pmatrix} 1 & 1 & \dots & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

X n'est pas de rang $p+1$ et donc $X^\top X$ est singulière. On impose en plus une contrainte d'équilibrage.

$$a_1 + a_2 + \dots + a_j + \dots + a_p = 0$$

Différents type de modèles linéaires (simples)

$$\mathbf{y} = X\mathbf{a} + \varepsilon$$

	variable à expliquer $\mathbf{y}$	variable explicative X	nom de la méthode
	quantitative	quantitative	régression
exemple :	position	temps	
	quantitative	qualitative	analyse de la variance
exemple :	rendement	variété	
	qualitative	quantitative	analyse discriminante
exemple :	variété	taille	
	qualitative	qualitative	analyse discriminante
exemple :	spam/nospam	présence de mots	

Conclusion

- le modèle linéaire
  - $\mathbf{y} = X\widehat{\alpha} + \varepsilon$
- Moindres carré
  - le cout est minimal quand $X^\top(\mathbf{y} - X\widehat{\alpha}) = 0$
  - calcul de $\widehat{\alpha}$ au prix de la résolution d'un système linéaire
- reste à faire
  - diagnostic du modèle
  - diagnostic des variables
  - diagnostic des individus
  - transformation des variables

Repéres bibliographiques

-