M1 IAD
Notes du cours RFIDEC (3)

Jean-Yves Jaffray

9 novembre 2006

1 Le modèle linéaire

1.1 La régression simple

Dans le chapitre de statistique descriptive, nous avons déjà abordé l'étude de la liaison entre deux variables quantitatives X et Y dont on possède n observations (x_i,y_i) et introduit un indicateur de liaison, le *coefficient de corrélation linéaire*,

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

Nous reprenons ici cette étude dans un cadre probabiliste. De plus, nous nous plaçons dans une problématique où les deux variables jouent des rôles dissymétriques :

i) Il y a une *variable expliquée* (dite encore *variable endogène*), Y, et une *variable explicative* (dite encore *variable exogène*), X : le modèle veut "expliquer" la valeur de Y par celle X.

Exemples

- *Conducteurs automobiles* : X, taux d'alcool dans le sang ; Y, vitesse.
- *Appartements* : X, surface du logement ; Y, prix au m^2.
- *Culture du blé* : X, quantité d'engrais à l'hectare ; Y, rendement à l'hectare.

ii) *La variable exogène X peut être aléatoire, mais pas nécessairement* ; dans le cas de la culture du blé, l'expérimentateur peut faire varier comme il l'entend la quantité d'engrais de parcelle en parcelle.

En revanche, on postule que, s'il y a une relation imprécise entre la valeur de X et celle de Y, c'est parce que cette dernière dépend aussi d'un deuxième facteur, aléatoire celui-là, $\mathcal{E}$, appelé *résidu* ou *erreur* (ou, dans certains contextes, *bruit*),

$$Y = f(X,\mathcal{E}).$$

L'existence de ce résidu peut venir de ce que f n'est pas la fonction exacte liant Y à X (en fait $Y = g(X)$, mais f est plus simple que g) ou de ce que Y ne dépend pas que de X (en fait $Y = F(X,X',X'',\cdots)$ mais on n'explicite pas cette relation).

Quoi qu'il en soit, $\mathcal{E}$ est supposé aléatoire, ce qui fait que *la variable endogène Y est elle-même une variable aléatoire.*

Dans le *modèle linéaire*, la fonction f est affine (= linéaire+cte)

$$Y = \alpha + \beta\, X + \mathcal{E},$$

mais α et β sont des paramètres inconnus.

On disposera de n observations (x_i,y_i) du couple (X,Y) . Elles seront liées par des relations

$$y_i = \alpha + \beta\, x_i + \epsilon_i.$$

L'existence des résidus ϵ_i fait que les points de coordonnées (x_i,y_i) ne sont pas portés par une même droite et que l'on ne pourra pas déterminer les valeurs exactes de α et β, mais seulement les estimer.

Remarque Le modèle linéaire est moins restrictif qu'il ne paraît. En effet, on peut jouer sur les changements de variables. Par exemple la relation $\ln Y = a + bX^2$ devient affine en prenant pour variables $Y' = \ln Y$ et $X' = X^2$.

1.1.1 Les hypothèses du modèle

$\mathcal{E}$ est une variable aléatoire et le n-uple $(\epsilon_1, \cdots, \epsilon_i, \cdots, \epsilon_n)$ est constitué de n tirages indépendants selon cette loi : c'est la réalisation (non observée) d'un n-échantillon $(\mathcal{E}_1, \cdots, \mathcal{E}_i, \cdots, \mathcal{E}_n)$ tiré de $\mathcal{E}$.

On supposera toujours que l'espérance de $\mathcal{E}$ est nulle :

$$E(\mathcal{E}) = 0$$

ce qui n'est pas restrictif (car on peut jouer sur la valeur de α) ; sa variance $V(\mathcal{E}) = \sigma^2$ sera en général inconnue ; c'est une hypothèse forte de supposer que tous les résidus ont même loi (par exemple, il pourrait se faire que l'écart-type du résidu, σ_i, soit proportionnel à x_i).

Certains résultats ne seront valables que sous l'*hypothèse de normalité des résidus* : $\mathcal{E}$ suit la loi $\mathcal{N}(0,\sigma)$.

Les x_i sont supposés être des constantes (même s'ils étaient aléatoires - variables X_i - on pourrait toujours examiner ce qu'on peut dire conditionnellement à $X_i = x_i$).

Chaque variable Y_i, de réalisation y_i, est alors aléatoire comme fonction d'une variable aléatoire $\mathcal{E}_i$ de même loi que $\mathcal{E}$:

$$Y_i = \alpha + \beta\, x_i + \mathcal{E}_i.$$

On peut en déduire, entre autres, que

$$E(Y_i) = \alpha + \beta\, x_i \text{ et } V(Y_i) = \sigma^2$$

et que, si $\mathcal{E}$ suit la loi $\mathcal{N}(0,\sigma)$, alors la loi de Y_i est $\mathcal{N}(\alpha + \beta\, x_i,\sigma)$.

1.1.2 La droite des moindres carrés

Dans un cadre purement descriptif, on peut associer aux points observés (x_i,y_i) la droite $y = a+bx$ dont ils sont le plus proches au sens précis suivant : la somme des carrés des distances (euclidiennes) verticales entre les points et la droite est la plus petite possible.
L'écart vertical entre le i^{eme} point et la droite est

$$e_i = y_i - (a + bx_i)\,;$$

on cherche donc

$$\min_{a,b} \sum_{i=1}^n e_i^2, \text{ c-à-d } \min_{a,b} \sum_{i=1}^n [y_i - a - bx_i]^2 = F(a,b)$$

Les conditions d'optimalité du premier ordre, qui sont ici conditions suffisantes d'optimalité car la fonction à minimiser est une fonction convexe de a et b sont :

$$\frac{\partial F(a,b)}{\partial a} = \sum_{i=1}^n (-2)[y_i - a - bx_i] = 0 \ (1)$$
$$\frac{\partial F(a,b)}{\partial b} = \sum_{i=1}^n (-2)x_i[y_i - a - bx_i] = 0 \ (2)$$

On voit que

$$(1) \Leftrightarrow a = \bar{y} - b\bar{x}\,;\ (2) \Leftrightarrow b\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - a\sum_{i=1}^n x_i$$

on peut éliminer a de (2) :

$$(1)+(2) \Rightarrow b\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i + nb(\sum_{i=1}^n x_i)^2$$
$$\Rightarrow b = \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i^2 - n(\sum_i x_i)^2},$$

d'où en divisant numérateur et dénominateur par n,

$$(1) \Leftrightarrow a = \bar{y} - b\bar{x}\,;\ (2) \Leftrightarrow b = \frac{cov(x,y)}{s_x^2}\,;$$

b est appelé *coefficient de régression (linéaire)* de Y en X.

On peut noter que, d'après (1), la droite des moindres carrés passe par le centre de gravité $(\bar{x},\bar{y})$ du nuage de points.

Propriétés des résidus des moindres carrés

La somme des résidus est nulle, car

$$e_i = y_i - (a + bx_i) \Rightarrow \sum_{i=1}^n e_i = n\bar{y} - na - bn\bar{x} = 0\,;$$

les résidus ne sont donc pas indépendants.

Les résidus ne sont pas corrélés aux valeurs de x car

$cov(e_i,x_i) = \frac{1}{n}\sum_{i=1}^n e_i[x_i - \bar{x}] = \frac{1}{n}\sum_{i=1}^n [y_i - (a+bx_i)][x_i - \bar{x}] = cov(y,x) - bs_x^2 = 0$

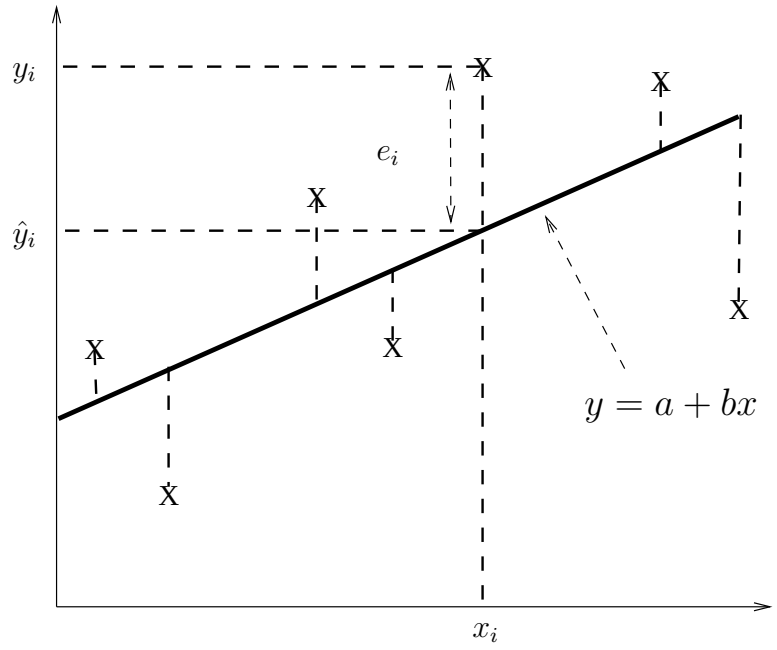

FIG. 1 – *Droite des moindres carrés*

Analyse de la variance

Posons $\hat{y}_i = a + bx_i$

On peut décomposer la variance empirique de Y, s_y^2 comme suit :

$$s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Le premier terme est la *variance expliquée* par le modèle ; le second, qui n'est autre que $\frac{1}{n}\sum_{i=1}^n e_i^2$ est la *variance résiduelle*. On pose

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

Le modèle linéaire rend d'autant mieux compte de la liaison entre X et Y que R^2 est plus proche de 1.

1.1.3 Propriétés des estimateurs des moindres carrés

Revenons au modèle linéaire probabiliste

$$Y = \alpha + \beta X + \mathcal{E},$$

et étudions les propriétés des solutions des moindres carrés, a et b en tant qu'estimateurs des paramètres α et β, respectivement.

Commençons par b ; d'après son expression, b est une réalisation de la variable B, aléatoire comme fonction (en fait combinaison linéaire) des variables aléatoires Y et $\bar{Y}$, donnée par

$$B = \frac{1}{\sum_i (x_i - \bar{x})^2} \times \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$$

(rappelons que les x_i sont des constantes).
Calculons l'espérance de B pour une valeur donnée (α,β) des paramètres :
Nous avons vu que $\forall i, E_{\alpha,\beta}(Y_i) = \alpha + \beta\, x_i$; on en déduit que :
$E_{\alpha,\beta}(\bar{Y}) = \frac{1}{n} \sum_i E_{\alpha,\beta}(Y_i) = \alpha + \beta\bar{x}$
et donc que
$E_{\alpha,\beta}(Y_i - \bar{Y}) = \beta(x_i - \bar{x})$ et, finalement
$E_{\alpha,\beta}(B) = \frac{1}{\sum_i (x_i - \bar{x})^2} \times \sum_i (x_i - \bar{x}) E_{\alpha,\beta}(Y_i - \bar{Y}) = \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \beta = \beta.$

Passons maintenant à a ; c'est une réalisation de la variable A, aléatoire comme fonction des variables aléatoires $\bar{Y}$ et B, puisque $A = \bar{Y} - B\bar{x}$.

On en déduit que

$$E_{\alpha,\beta}(A) = E_{\alpha,\beta}\bar{Y} - E_{\alpha,\beta}(B)\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha.$$

a et b sont donc des estimateurs sans biais des paramètres α et β,respectivement.

Etudions maintenant leurs variances $\sigma_B^2 = V_{\alpha,\beta}(B)$ et $\sigma_A^2 = V_{\alpha,\beta}(A)$ (en fait, elles ne dépendront pas de α ni de β).

Partant de l'expression $cov(x,y) = \sum_i (x_i - \bar{x}) y_i$, nous pouvons écrire

$$B = \frac{1}{n s_x^2} \sum_i (x_i - \bar{x}) Y_i.$$

Les variables Y_i étant mutuellement indépendantes (parce que les $\mathcal{E}_i$ le sont), la variance de B vaut

$$\sigma_B^2 = \frac{1}{n^2 s_x^4} \sum_i (x_i - \bar{x})^2 V(Y_i) = \frac{n s_x^2}{n^2 s_x^4} \sigma^2 \Leftrightarrow \sigma_B^2 = \frac{\sigma^2}{n s_x^2}$$

Par ailleurs, de $A = \bar{Y} - \bar{x}B$, on déduit que

$$\sigma_A^2 = V(\bar{Y}) + (\bar{x})^2 \sigma_B^2 - 2\bar{x} cov(\bar{Y},B) \text{ ; or}$$
$$cov(\bar{Y},B) = \frac{1}{n s_x^2} \sum_i (x_i - \bar{x}) cov(\bar{Y},Y_i) = \frac{1}{n s_x^2} \sum_i (x_i - \bar{x}) \frac{\sigma^2}{n} =$$
$$\frac{\sigma^2}{s_x^2} \sum_i (x_i - \bar{x}) = 0,$$

puisque, par indépendance de Y_i et Y_j pour $i \neq j$,

$$cov(\bar{Y},Y_i) = \frac{1}{n} V(Y_i) = \frac{\sigma^2}{n} ;$$

finalement,

$$\sigma_A^2 = V(\bar{Y}) + (\bar{x})^2 \sigma_B^2 = \frac{\sigma^2}{n} (1 + \frac{(\bar{x})^2}{s_x^2})$$

a et b sont donc des estimateurs convergents des paramètres α et β.
Ils sont en fait d'efficacité maximale dans une certaine classe d'estimateurs, comme le montre le résultat suivant que nous admettrons :

Théorème de GAUSS-MARKOV. A et B sont de variance minimale parmi les estimateurs sans biais de α et β qui sont fonctions linéaires des Y_i.

On peut aussi avoir à estimer σ^2 qui est inconnu. La somme des résidus $\sum_i e_i^2$ devrait être d'autant plus grande que σ a une valeur plus élevée et, en effet, on démontre que la variable $\hat{\Sigma}^2$ de valeurs:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i e_i^2 = \frac{1}{n-2}\sum_i [y_i - (a + bx_i)]^2$$

est un estimateur sans biais de σ^2.

1.1.4 Test de signification du coefficient de régression linéaire b

La droite des moindres carrés existe toujours ; on peut donc s'interroger sur l'existence effective de la liaison linéaire postulée entre X et Y.
Pour cela on va tester

$H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$

On a vu que

$$E_{\alpha,\beta}(B) = \beta \text{ et } \sigma_B^2 = \frac{\sigma^2}{ns_x^2}$$

σ^2 étant inconnu, σ_B^2 l'est aussi ; on prend pour estimateur $\hat{\Sigma}_B^2$ de valeurs

$$\hat{\sigma}_B^2 = \frac{\hat{\sigma}^2}{ns_x^2}.$$

Sous l'hypothèse de normalité des résidus, $\frac{B-\beta}{\hat{\Sigma}_B}$ suit alors, dans l'hypothèse H_1, la loi de STUDENT à $n-2$ degré de liberté ; dans l'hypothèse H_0, c'est simplement $\frac{B}{\hat{\Sigma}_B}$ qui suit cette loi.

Un test à niveau signification η, c-à-d une probabilité d'erreur de première espèce (rejeter à tort H_0) égale à η, peut être obtenu ainsi :

Soit t (lu dans la table) tel que $P(\mid T_{n-2} \mid > t) = \eta$
(T_{n-2} variable de loi de STUDENT à $n-2$ d.l.)

$$\text{Accepter } H_0 \Leftrightarrow \mid \frac{b}{\hat{\sigma}_B} \mid \leq t.$$

L'intervalle d'acceptation de H_0 est tel que les probabilités, sous H_0, de sortir de l'intervalle à gauche et à droite soient égales (chacune vaut $\frac{\eta}{2}$).

1.1.5 Intervalle de confiance pour le coefficient de régression linéaire b

Toujours *sous l'hypothèse de normalité des résidus*, en s'appuyant sur le fait que $\frac{B-\beta}{\hat{\Sigma}_B}$ suit une loi de STUDENT à $n-2$ degré de liberté, l'intervalle

$$[b - t\hat{\sigma}_B, b + t\hat{\sigma}_B] \ ,$$

où t est tel que $P(\mid T_{n-2} \mid > t) = \eta$,
est un intervalle de confiance à $100(1-\eta)\%$.

1.1.6 Prévision dans le modèle linéaire

Estimation

Les coefficients a et b étant estimés à partir des observations, on peut prédire la valeur que prendrait Y si la valeur de X était x_0.
On prend pour estimation de Y l'ordonnée du point de la droite des moindres carrés d'abscisse x_0, c-à-d $y_0^* = a + bx_0$.
Autrement dit, on utilise pour estimateur de la valeur vraie (inconnue) $Y_0 = \alpha + \beta x_0$ de Y la variable aléatoire $Y_0^* = A + Bx_0$.
Son espérance est
$E_{\alpha,\beta}[Y_0^*] = E_{\alpha,\beta}(A) + E_{\alpha,\beta}(B)x_0 = \alpha + \beta x_0$: *c'est donc un estimateur sans biais de la valeur vraie* Y_0 *de* Y lorsque $X = x_0$

Comme $Y_0^* = A + Bx_0 = \bar{Y} - \bar{x}B + x_0B = \bar{Y} + (x_0 - \bar{x})B$ et que $cov(\bar{Y},B) = 0$, sa variance vaut

$$V(Y_0^*) = V(\bar{Y}) + (x_0 - \bar{x})^2\sigma_B^2 = \frac{\sigma^2}{n}[1 + (x_0 - \bar{x})^2\frac{\sigma_B^2}{\sigma^2/n}] = \frac{\sigma^2}{n}[1 + \frac{(x_0-\bar{x})^2}{s_x^2}]$$

Comme σ^2 est en général inconnu et estimé par $\hat{\sigma}^2$, $V(Y_0^*)$ sera estimé par $\frac{\hat{\sigma}^2}{n}[1 + \frac{(x_0-\bar{x})^2}{s_x^2}]$.

Intervalle de confiance

Sous l'hypothèse de normalité des résidus,

$$\frac{Y_0 - Y_0^*}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}}$$

suit une loi de STUDENT à $n-2$ degrés de liberté.

L'intervalle

$$[Y_0^* - t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}, Y_0^* + t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}]$$

où t est tel que $P(\mid T_{n-2} \mid > t = \eta$,
aura $100(1-\eta)\%$ de chances de contenir Y_0.

1.2 La régression multiple

Le modèle de régression simple se généralise de manière naturelle au cas où la variable endogène, Y, doit être expliquée par plusieurs variables exogènes, $X_j, j = 1, \cdots, p$.

La relation postulée est

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_p X_p + \mathcal{E},$$

où les paramètres β_j sont inconnus.

On dispose de n observations du (p+1)-uple $(x_1,\cdots,x_j,\cdots,x_p,y)$ du couple (X,Y) ; les composantes de la i^{eme} observation sont donc reliées par la relation :

$$y_i = \beta_0 + \beta_1\, x_{i,1} + \cdots + \beta_j\, x_{i,j} + \cdots + \beta_p\, x_{i,p} + \epsilon_i.$$

Il est commode d'utiliser une notation matricielle :

$$Y_{n\times 1} = [y_i]\,;\; X_{n\times(p+1)} = [x_{i,j}]\,;\beta_{(n+1)\times 1} = [\beta_j]\,;\epsilon_{n\times 1} = [\epsilon_i]$$

[la première colonne de la matrice X est le vecteur de composantes toutes égales à 1]
d'où

$$Y \;=\; X\,\beta \,+\, \epsilon$$

Les hypothèses sont les mêmes que dans le cas de la régression simple : les variables X_j sont observées exactement et sont donc des constantes ; le n-uple $(\epsilon_1,\cdots,\epsilon_i,\cdots,\epsilon_n)$ est constitué de n tirages indépendants de la même variable $\mathcal{E}$ d'espérance nulle et de variance σ^2.
Y est donc aléatoire comme fonction de $\mathcal{E}$.

La recherche de

$$\min_\beta \textstyle\sum_{i=1}^n e_i^2,$$

$$\text{c-à-d } \min_{(\beta_j,j=0,..,p)} \textstyle\sum_{i=1}^n [y_i - \sum_{j=0}^p \beta_j\, x_{i,j}]^2$$

conduit à l'estimateur des moindres carrés du vecteur de paramètres β :

$$\hat{\beta} = ({}^\tau XX)^{-1}.{}^\tau XY$$

où ${}^\tau X$ est la transposée de la matrice X.

On démontre que $\hat{\beta}$ est un estimateur sans biais de β et que sa matrice des variances-covariances est $V(\hat{\beta}) = ({}^\tau XX)^{-1}\sigma^2$.

Remarque. $({}^\tau XX)^{-1}$ peut ne pas exister ; c'est en particulier le cas lorsque une des variables explicatives est combinaison linéaire de certaines autres.