I. Introduction aux modèles stochastiques

1 Exemple introductif

Considérons les données portées au tableau ci-dessous, extraites d'une population d'entreprises ; leur part de marché X ainsi que leur rentabilité Y, au cours de la même période, sont exprimées en pourcentage.

Part de marché et rentabilité

Observation	**Part de marché *X*** **(en %)**	**Rentabilité *Y*** **(en %)**	$p(Y_i/X_i)$	$E(Y_i/X_i)$
1	10	3	¼	4
2	10	4	½	
3	10	4		
4	10	5	¼	
5	20	5	1/3	6
6	20	6	1/3	
7	20	7	1/3	
8	30	7	1/3	8
9	30	8	1/3	
10	30	9	1/3	
…	…	…	…	…

La population de taille N a été partagée en *n* sous-groupes selon les valeurs de X. On dispose ainsi de *n* valeurs de X et des valeurs de Y associées à chacune, soit *n* sous-populations de Y et *n* moyennes conditionnelles de Y sachant X. Les trois moyennes conditionnelles de Y associées aux trois valeurs de X dans le tableau ont été calculées, sachant que :

$$E(Y_i/X_i) = \sum_{i=1}^{i=k} Y_i \; p(Y_i/X_i)$$

Pour chacune des sous-populations de Y, on observe une variation des rentabilités, autour des moyennes conditionnelles de Y, mais en dépit de ces fluctuations, en moyenne, Y croît avec l'augmentation de X.

Chaque valeur de Y trouve une expression à partir de l'équation de régression :

$$Y_i = E(Y/X_i) + \varepsilon_i \text{ ou } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Ainsi, pour X = 10,

$$Y_1 = 3 = \beta_0 + \beta_1 .10 + \varepsilon_1$$
$$Y_2 = 4 = \beta_0 + \beta_1 .10 + \varepsilon_2$$
$$Y_3 = 4 = \beta_0 + \beta_1 .10 + \varepsilon_3$$
$$Y_4 = 5 = \beta_0 + \beta_1 .10 + \varepsilon_4$$

Puisque $E(Y/10) = \beta_0 + \beta_1 .10 = 4$,

$$\varepsilon_1 = -1 \; ; \; \varepsilon_2 = 0 \; ; \; \varepsilon_3 = 0 \; ; \; \varepsilon_4 = +1$$

$$\text{et } E(\varepsilon_i/10) = 0$$

Dans le cas général,

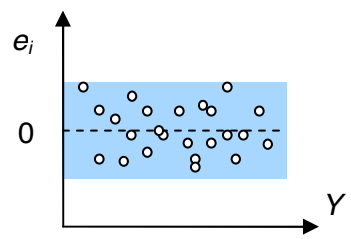
$$E(\varepsilon_i/X_i) = 0$$

ce qui revient à poser que les moyennes conditionnelles des résidus ont une valeur nulle. Cette hypothèse est fondamentale aux modèles de régression.

2 À propos des résidus

Lorsque les résidus sont indépendants, ont une espérance nulle et une variance constante, leur représentation graphique en fonction des valeurs de *Y* est la suivante :

Forme correcte des résidus



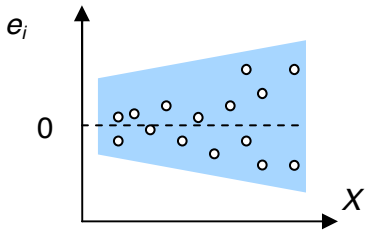
La variance des termes d'erreurs est constante. Les résidus gravitent autour de zéro. On pourrait également représenter le comportement des résidus « standardisés », en divisant chaque résidu par son écart type, et accepter l'hypothèse de normalité des termes d'erreurs si 95% des résidus « standardisés » sont compris dans l'intervalle [-2 ; 2].

Si le modèle comporte plusieurs régresseurs, comme dans le cas de la régression multiple, on prendra le soin de réaliser une représentation graphique des résidus par variable intégrée au modèle et vérifier que les résidus ont bien une forme horizontale pour chacune, afin de ne pas être conduit à remettre en question les hypothèses sur lesquelles repose le modèle de régression.

Si des points extrêmes sont isolés du reste du nuage des résidus, les données sont à vérifier. S'il n'y a pas d'erreur de saisie, l'utilisation d'une technique moins sensible que les moindres carrés aux points « aberrants » s'impose.

D'autres cas peuvent se présenter. Si la variance de *e* n'est plus constante, mais disons croissante, c'est-à-dire plus élevée pour les valeurs de X les plus grandes, les résidus épouseront une forme similaire à celle de la figure suivante.

Forme croissante des résidus



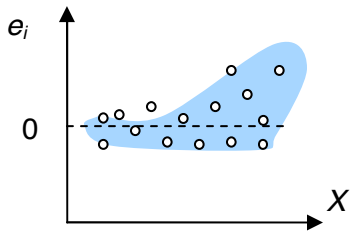
En ce cas, l'hypothèse d'une variance constante des résidus n'est plus valide :

$$E\left[e_i - E(e_i)\right]^2 \neq \sigma^2(\varepsilon)$$

Dans cette situation, il y a hétéroscédasticité de la variance. Certaines formes apportent une information sur l'origine de l'hétéroscédasticité. Des résidus en forme d'arc, avec une dispersion croissante, comme dans le cas de la figure ci-dessous, signalent qu'il faut procéder à une transformation logarithmique de *Y* et régresser plutôt Log Y en X, en posant :

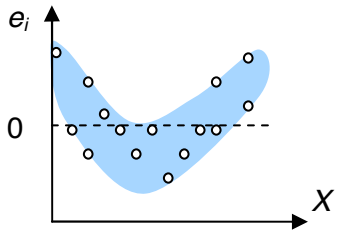
$$\text{Log } \hat{Y}_i = b_0 + b_1 X_i + e_i$$

Forme nécessitant une régression Log Y

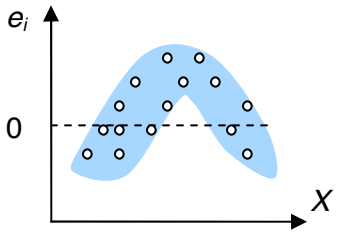


Si les résidus ont une forme en arc, plus symétrique, la relation qui unit *X* à *Y* n'est pas linéaire mais curviligne. La répartition des points suggère l'adoption d'une forme non linéaire. Le modèle nécessite ici l'introduction d'un terme au carré.

Formes impliquant un modèle de second ordre



(a) Forme en U



(b) Forme en ∩

Il s'agit de poser alors une fonction de régression sur l'échantillon du type :

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 + e_i$$

La liaison n'est plus une relation fonctionnelle simple, il n'y a plus de droite qui passe approximativement par tous les points. Ce type de modèle est appelé modèle de régression « quadratique ». En plus du problème de l'hétéroscédasticité, un autre problème concerne l'autocorrélation des résidus.

II. Le modèle de régression linéaire simple

1 La fonction de régression affine

La fonction de régression affine $\hat{Y}_i = b_0 + b_1 X_i + e_i$ a pour domaine de définition l'ensemble des nombres réels (il n'y a en effet ni dénominateur, ni racine carrée, ni logarithme dans le calcul de Y). La fonction étant une droite, elle est sa propre asymptote (ainsi quand X tend vers $\pm \infty$, Y tend vers $\pm \infty$). Enfin, la dérivée première de $\hat{Y}$ par rapport à X est égale à :

$$\frac{\partial \hat{Y}}{\partial X} = b_1$$

La fonction est ainsi croissante sur son domaine de définition si $b_1 > 0$ et décroissante si $b_1 < 0$. En revanche, la dérivée seconde :

$$\frac{\partial^2 \hat{Y}}{\partial^2 X} = 0$$

Pour trouver le modèle qui approche le mieux la dépendance entre *X* et *Y*, il faut déterminer les valeurs des paramètres b_0 et b_1 de la droite qui minimise la distance entre les valeurs de *Y*, réellement observées, et les valeurs de $\hat{Y}$, calculées par application de l'équation de régression.

2 Formulation des estimateurs

Mathématiquement, c'est en posant égales à zéro les dérivées partielles de la fonction que l'on peut déterminer les valeurs des estimateurs :

$$\frac{\partial \sum e_i^2}{\partial b_0} = 2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)(-1) = -2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right) \equiv 0$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = 2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)(-X_i) = -2 \sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i\right) \equiv 0$$

En cherchant à annuler chacune des deux dérivées partielles, il vient :

$$\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right) = 0 \text{ et } \sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i\right) = 0$$

Les estimateurs b_0 et b_1 sont les solutions de ce système d'équations. En développant,

$$\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right) = 0$$

on obtient tout d'abord :

$$\sum Y_i - \sum b_0 - b_1 \sum X_i = 0$$

$$\sum Y_i - n\, b_0 - b_1 \sum X_i = 0$$

$$b_0 = \frac{\sum Y_i - b_1 \sum X_i}{n} = \overline{Y} - b_1 \overline{X}$$

À partir de cette formule, l'on peut écrire :

$$\overline{Y} = b_0 + b_1 \overline{X}$$

La droite de régression de l'échantillon passe par donc par le couple de coordonnées $\left(\overline{X},\overline{Y}\right)$. On trouve la valeur de l'estimateur b_1 en solutionnant :

$$\sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i\right) \equiv 0$$

On a :

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

et

$$b_1 \sum X_i^2 = \sum X_i Y_i - b_0 \sum X_i \text{ où } b_0 \sum X_i = b_0\, n\, \overline{X}$$

L'égalité devient :

$$b_1 \sum X_i^2 = \sum X_i Y_i - \left(\overline{Y} - b_1 \overline{X}\right) n\, \overline{X}$$

$$b_1 \sum X_i^2 = \sum X_i Y_i - n\, \overline{X}\, \overline{Y} + b_1\, n\, \overline{X}^2$$

d'où,

$$b_1 \sum X_i^2 - b_1 n \overline{X}^2 = \sum X_i Y_i - n \overline{X}\ \overline{Y}$$
$$b_1 \left(\sum X_i^2 - n \overline{X}^2 \right) = \sum X_i Y_i - n \overline{X}\ \overline{Y}$$

Par conséquent :

$$b_1 = \frac{\sum X_i Y_i - n \overline{X}\ \overline{Y}}{\sum X_i^2 - n \overline{X}^2}$$

Le couple (b_0, b_1) est bien un minimum car les dérivées secondes de la fonction $\sum e_i^2$ sont positives. Les dérivées premières de la fonction donnaient :

$$\frac{\partial \sum e_i^2}{\partial b_0} = -2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i \right)$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2 \sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i \right)$$

En développant, il vient :

$$\frac{\partial \sum e_i^2}{\partial b_0} = -2 \sum Y_i + 2 n b_0 + 2 b_1 \sum X_i$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2 \sum X_i Y_i + 2 b_0 \sum X_i + 2 b_1 \sum X_i^2$$

avec :

$$\frac{\partial^2 \sum_{i=1}^{n} e_i^2}{\partial b_0^2} = 2n > 0 \; ; \frac{\partial^2 \sum_{i=1}^{n} e_i^2}{\partial b_1^2} = 2 \sum_{i=1}^{n} X_i^2 > 0 \; ;$$

$$\frac{\partial^2 \sum_{i=1}^{n} e_i^2}{\partial b_0\ \partial b_1} = 2 \sum_{i=1}^{n} X_i > 0$$

En pratique, on utilise une formule qui réduit le nombre des calculs pour déterminer b_1 :

$$b_1 = \frac{\sum \left(X_i - \overline{X} \right) Y_i}{\sum \left(X_i - \overline{X} \right)^2}$$

On utilise également régulièrement une troisième formule pour obtenir une estimation ponctuelle de l'estimateur :

$$b_1 = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

Les deux dernières équations signalent que l'estimation de la pente de la droite de régression est soumise à l'existence d'une certaine variabilité de X. Si la variance de X est nulle, le coefficient b_1 ne peut pas être calculé.

Ainsi, la pente b_1, qui accompagne une variation unitaire de X peut être obtenue à partir de trois formules équivalentes :

$$b_1 = \frac{\sum X_i Y_i - n\,\overline{X}\,\overline{Y}}{\sum X_i^2 - n\,\overline{X}^2} \; ; \; b_1 = \frac{\sum \left(X_i - \overline{X}\right) Y_i}{\sum \left(X_i - \overline{X}\right)^2} \; ;$$

$$b_1 = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

tandis que b_0, l'ordonnée à l'origine, est obtenue à partir de la différence suivante :

$$b_0 = \overline{Y} - b_1\,\overline{X}$$

III. Approche matricielle de la régression : une application

1 Énoncé

Les données dans le tableau ci-dessous concernent 10 entreprises de l'industrie chimique. On cherche à établir une relation entre la production Y, les heures de travail X_1 et le capital utilisé X_2.

Données de l'échantillon

Entreprise	Production Y (100 tonnes)	Travail X_1 (heures)	Capital X_2 (machines/heures)
1	60	1100	300
2	120	1200	400
3	90	1430	420
4	250	1500	400
5	300	1520	510
6	360	1620	590
7	380	1800	600
8	430	1820	630
9	440	1800	610
10	490	1750	630

1) Calculer l'approximation linéaire de Y en X_1 en utilisant l'approche matricielle à l'aide d'un ordinateur.

2) On fait à présent l'hypothèse d'un modèle de régression multiple avec deux variables explicatives X_1 et X_2. Déterminer le vecteur des estimateurs. Quelles valeurs obtient-on pour les paramètres de l'équation de régression avec le critère des moindres carrés ?

3) Le modèle de régression multiple vous semble-t-il satisfaisant ?

2 Solution

1) On obtient les paramètres de la droite de régression simple grâce à la méthode des moindres carrés en calculant le vecteur :

$$b = \left(X' X\right)^{-1} X' Y$$

$$X' X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1100 & 1200 & 1430 & 1500 & 1520 & 1620 & 1800 & 1820 & 1800 & 1750 \end{bmatrix} \begin{bmatrix} 1 & 1100 \\ 1 & 1200 \\ 1 & 1430 \\ 1 & 1500 \\ 1 & 1520 \\ 1 & 1620 \\ 1 & 1800 \\ 1 & 1820 \\ 1 & 1800 \\ 1 & 1750 \end{bmatrix}$$

$$X' X = \begin{bmatrix} 10 & 15\,540 \\ 15\,540 & 24\,734\,600 \end{bmatrix}$$

$$\left(X' X\right)^{-1} = \begin{bmatrix} 4{,}224959 & -0{,}002654 \\ -0{,}002654 & 0{,}000002 \end{bmatrix}$$

$$X' Y = \begin{bmatrix} 2\,920 \\ 4\,869\,000 \end{bmatrix}$$

$$b = \left(X' X\right)^{-1} X' Y = \begin{bmatrix} -587{,}460 \\ 0{,}566 \end{bmatrix}$$

La droite de régression de Y en X_1 a pour équation :

$$\hat{Y} = -587{,}460 + 0{,}566\ X_1$$

2) On obtient les paramètres de la droite de régression multiple de pareille façon grâce à la méthode des moindres carrés en calculant le vecteur des estimateurs :

$$b = \left(X' X\right)^{-1} X' Y$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1100 & 1200 & 1430 & 1500 & 1520 & 1620 & 1800 & 1820 & 1800 & 1750 \\ 300 & 400 & 420 & 400 & 510 & 590 & 600 & 630 & 610 & 630 \end{bmatrix} \begin{bmatrix} 1 & 1100 & 300 \\ 1 & 1200 & 400 \\ 1 & 1430 & 420 \\ 1 & 1500 & 400 \\ 1 & 1520 & 510 \\ 1 & 1620 & 590 \\ 1 & 1800 & 600 \\ 1 & 1820 & 630 \\ 1 & 1800 & 610 \\ 1 & 1750 & 630 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 10 & 15\,540 & 5\,090 \\ 15\,540 & 24\,734\,600 & 8\,168\,700 \\ 5\,090 & 8\,168\,700 & 2\,720\,500 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 6{,}304288 & -0{,}007817 & 0{,}011677 \\ -0{,}007817 & 0{,}000015 & -0{,}000029 \\ 0{,}011677 & -0{,}000029 & 0{,}000066 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 2\,920 \\ 4\,869\,000 \\ 1\,645\,200 \end{bmatrix}$$

$$b = (X'X)^{-1} X'Y = \begin{bmatrix} -442{,}269 \\ 0{,}205 \\ 0{,}815 \end{bmatrix}$$

La droite de régression de Y en X_1 a pour équation :

$$\hat{Y} = -442{,}269 + 0{,}205\ X_1 + 0{,}815\ X_2$$

3) Pour que le modèle de régression multiple soit satisfaisant, il faut que les variables X_i soient indépendantes. L'existence de corrélations significatives entre les variables explicatives suffit à rendre instable l'estimation des coefficients de régression. On obtient un indice de colinéarité en calculant ainsi le coefficient de corrélation entre les variables explicatives directement sous Excel grâce à la fonction « COEFFICIENT.CORRELATION » :

$$r_{X_1X_2} = 0{,}939$$

La présence de colinéarité entre les variables est forte. De fait, les coefficients du modèle sont biaisés. Notons toutefois que le capital utilisé X_2 et les heures de travail X_1 sont très fortement corrélés à la production :

$$r_{Y X_1} = 0{,}926$$

$$r_{Y X_2} = 0{,}942$$

De ce point de vue, les variables explicatives sont en apparence, en apparence seulement, quasiment substituables. Comment arbitrer alors entre X_1 et X_2 ? L'idéal serait de régresser Y en X_2 comme nous l'avons fait en X_1, et à la suite de comparer les résultats fournis par les tests. La variable la plus significativement liée à Y serait celle qui serait préservée.

Ce sont les valeurs de *t* ou les intervalles de confiance qui nous permettent de nous prononcer sur le maintien ou sur l'exclusion d'une variable explicative dans le modèle de régression multiple, en nous basant sur les écarts types des estimateurs. Ici, compte tenu du degré de corrélation entre X_1 et X_2, il faut s'attendre à une perte de significativité sur l'une d'entre elles