- Le plus souvent, un modèle éconmétrique devra faire appel à plusieurs variables explicatives, ce qui amènera à recourir à des techniques de *régression multiple*, et ce qui demandera également de résoudre les *problèmes pratiques* liés à la définition et à la sélection des variables.

1

a) La régression multiple:

- Un modèle de régression linéaire multiple vise à expliquer la valeur de la variable dépendante Y comme une combinaison linéaire des valeurs des variables explicatives X_1,…, X_n:

$$Y=a+b_1X_1+\ldots+b_nX_n$$

- Un tel modèle peut s'appliquer chaque fois que les relations entre Y et les X_i sont linéaires et chaque fois que des transformations appropriées permettent de retrouver le schéma linéaire.

2

- Par exemple, un modèle multiplicatif du type

$$Y = aX_1^{b_1} X_2^{b_2} \cdots X_n^{b_n}$$

devient linéaire en appliquant des logarithmes, et l'évolution des coefficients a, $b_1,\ldots,b_n$ est réalisée à l'aide de la régression linéaire multiple.

- Les coefficients optimaux sont déterminés par la méthode des moindres carrés. Dans le cas de deux variables explicatives X_1 et X_2, il s'agira de minimiser l'expression

$$E^2=\Sigma(Y-a-b_1X_1-b_2X_2)^2$$

par rapport à a, b_1 et b_2.

3

- Ces coefficients seront solution du système formé par les trois équations suivantes:

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1 Y = a\sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2$$

$$\sum X_2 Y = a\sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$$

• Les coefficients b_i ont une signification similaire à celle qui avait été notée dans le cas de la régression linéaire simple: ils indiquent de combien varie Y en moyenne pour une variation unitaire de X_i.

• Les mêmes tests de qualités et de fiabilité de la représentation se retrouvent également.

4

Exemple:

- Le tableau suivant donne les résultats d'une étude transversale sur dix régions: Les ventes Y du produit sont mises en relation avec les dépenses X_1 de publicité-presse et les dépenses X_2 de publicité sur les lieux de vente (PLV). (L'unité est 10^3 DH)

Observations (i)	Ventes (Y)	Publicité presse (X_1)	PLV (X_2)
1	30	2	6
2	22	1	3
3	29	6	2
4	35	4	5
5	25	3	3
6	40	2	8
7	24	6	1
8	21	2	2
9	32	7	2
10	15	1	1

5

- Les calculs présentés dans le tableau suivant conduisent à l'expression: $Y=2,15X_1+3,13X_2+9,65$

Observation	Y	X_1	X_2	X_1Y	X_2Y	X_1^2	X_2^2	X_1X_2
1	30	2	6	60	180	4	36	12
2	22	1	3	22	66	1	9	3
3	29	6	2	174	58	36	4	12
4	35	4	5	140	175	16	25	20
5	25	3	3	75	75	9	9	9
6	40	2	8	320	320	4	64	16
7	24	6	1	24	24	36	1	6
8	21	2	2	42	42	4	4	4
9	32	7	2	64	64	49	4	14
10	15	1	1	15	15	1	1	1
Total……	273	34	33	976	1019	160	157	97
Moyenne....	27,3	3,4	3,3					

6

- On a les équations suivantes:

$$\left.\begin{aligned} 273 &= 10\cdot a + 34b_1 + 33b_2 \\ 976 &= 34\cdot a + 160b_1 + 97b_2 \\ 1019 &= 33\cdot a + 97b_1 + 157b_2 \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} a &= 9{,}65 \\ b_1 &= 2{,}15 \\ b_2 &= 3{,}13. \end{aligned}\right.$$

- Ce modèle explique 96,8% ($R^2 = 0{,}968$) de la dispersion totale avec un coefficient de corrélation multiple $R = 0{,}983$.

7

- Le test F est également favorable:

	Degrés de liberté	Somme des carrés	Carrés moyens	F
Régression… Erreur…….	k = 2 n-k-1 = 7	472 15	236 2,14	110
Total	n-1 = 9	477		

$F_{0,01} = 9{,}55$

• De même que le test t sur chacun des coefficients de régression partielle b_1 et b_2:

- Écart-type des erreurs: $S_{Y\cdot 12} = \sqrt{\dfrac{\sum\left(\hat{Y} - Y\right)^2}{n-3}} = 1{,}47$

- Erreur-standard sur b_1: $S_{b_1} = \dfrac{S_{Y\cdot 12}}{\sqrt{\sum X_1^2 - nX_1^2}} = 0{,}23$

8

- Erreur-standard sur b2:

$$S_{b_2} = \frac{S_{Y\cdot 12}}{\sqrt{\sum X_2^2 - nX_2^2}} = 0,22$$

- Test t pour b_1:

$$t_1 = \frac{2,15}{0,22} = 9,3$$

- pour b_2:

$$t_2 = \frac{3,13}{0,21} = 14,2$$

$t_{0,01}$=2,9 pour n-k-1=7 degrés de liberté.

9

b) Problèmes pratiques:

- Dans le cadre de modèles exogènes, on peut être amené à traiter un grand nombre de variables explicatives, dont *a priori* le pouvoir explicatif n'est pas connu. Il serait certes concevable d'appliquer la technique décrite ci-dessus à toutes les combinaisons envisageables de variables explicatives. Mais cette procédure entraînerait de trop nombreux calculs. Certains programmes informatiques permettent à partir d'un ensemble de variables explicatives potentielles de ne retenir que celles qui sont les plus intéressantes, en introduisant succéssivement dans l'équation de régression les variables qui absorbent le plus de variance résiduelle.

10

- Quand plusieurs variables interviennent dans le modèle de régression, il faut envisager le risque de *multicolinéarité*. Ce phénomène se produit quand deux variables explicatives ou plus sont fortement corrélées entre elles. Une bonne définition des coefficients b_i demande de ne retenir que des variables indépendantes. Une première opération consiste donc, avant tout traitement par le modèle de régression, à examiner la matrice des coefficients de corrélation entre les X_i et à sélectionner pour chaque ensemble de variables corrélées un seul représentant.

11

- La nature des variables utilisables dans les modèles de régression peut être une source de difficulté. Les variables quantitatives sont les plus adaptées à ce type de formalisation. Des variables nominales (Les variables nominales correspondent à des catégories,comme par exemple, classes d'âge, de revenu, localisation géographique,…etc) peuvent également être utilisées, mais il faut alors recourir à des variables muettes (valeur égale à 0 ou 1) ce qui alourdit la formulation. Quand on ne dsipose que de variables nominales, il est préférable de s'en remettre à d'autres types de formulation comme la segmentation ou l'analyse de variance.

12