IV - REGRESSION LINEAIRE

J-P. Croisille

Université de Lorraine

UEL - Année 2012/2013

1- DEPENDANCE LINEAIRE

EQUATIONS LINEAIRES AVEC UNE VARIABLE INDEPENDANTE

Une variable **réponse** ou **sortie** est supposée être expliquée par des variables dites **explicatives**, appellées aussi **covariables**.
Une analyse de type *régression linéaire simple* est naturelle, lorsque dans une expérience, deux quantitées sont mesurées, et lorsque l'on souhaite expliquer la valeur d'une des deux variables par l'autre.

EXEMPLE DE DEPENDANCE LINEAIRE (1)

Une sociéte d'édition propose un service de traitement de texte à ses clients. Le tarif est 20 euros de l'heure auquel s'ajoute 25 euros forfaitaire de frais de stockage. Le coût y de x heures de travail est donc de:

$$y = 20x + 25 \tag{1}$$

L'équation $y = 20x + 25$ est linéaire. Elle s'écrit $y = ax + b$. On dit que

- a est la pente: c'est la valeur dont augmente y (la variable expliquée) quand x augmente de 1, (la variable explicative).
- b est l'ordonnée à l'origine. C'est la valeur de y là où la droite coupe l'axe des ordonnées.

EXEMPLE DE DEPENDANCE LINEAIRE (2)

Temps (en h)	Coût (en euros)
5.0	125
7.5	175
15.0	325
20.0	425
22.5	475

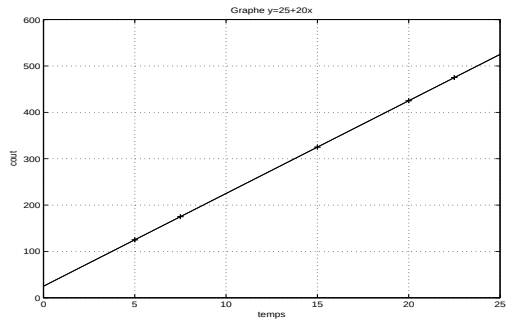

Figure: Fonction linéaire $x \mapsto y(x) = 20x + 25$

2- REGRESSION LINEAIRE SIMPLE

MODELES:

Dans la vie réelle, les exemples où on a une dépendance aussi rigoureuse (exactement linéaire) sont rares. Au contraire, on peut prédire de façon approchée la variable expliquée en fonction de la variable explicative, mais pas rigoureusement.

EXEMPLE:

Prédiction du prix d'une voiture d'occasion en fonction de son âge.
On collecte le tableau de données suivant (prix d'un modèle précis de RENAULT d'occasion)

Voiture	Age (en années)	Prix (en 100 euros)
1	5.0	85
2	4.0	103
3	6.0	70
4	5.0	82
5	5.0	89
6	5.0	98
7	6.0	66
8	6.0	95
9	2.0	169
10	7.0	70
11	7.0	48

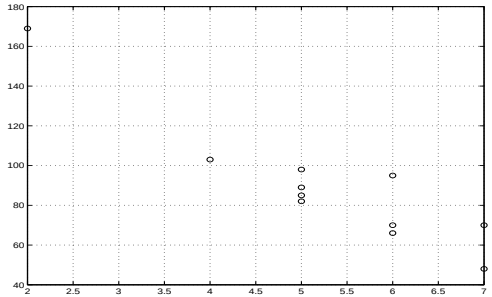

Figure: Abscisse: âge, Ordonnée: prix

DROITE DES MOINDRES CARRES

Soit $y = ax + b$ l'équation de la droite de paramètrs a et b. En chaque mesure point x_i, on a

- Une valeur mesurée y_i.
- Une valeur "prédite" $\hat{y}_i$.

. On appelle *erreur au point* x_i la quantité $e_i = \hat{y}_i - y_i$. La **droite des moindres carrés** est l'unique droite qui minimise la somme des carrés des erreurs, c'est-à-dire la quantité

$$\varphi(a, b) = \sum_{i=1}^{n} e_i^2 \tag{2}$$

L'équation $y = ax + b$ trouvée par ce critère s'appelle la droite de **régression**.

EQUATION DE LA DROITE DE REGRESSION

Les notations suivantes sont classiques:

- $$S_{xx} = \sum_i (x_i - \bar{x})^2 \quad = \sum_i x_i^2 - (\sum_i x_i)^2/n \tag{3}$$

- $$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad = \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)/n \tag{4}$$

- $$S_{yy} = \sum_i (y_i - \bar{y})^2 \quad = \sum_i y_i^2 - (\sum_i y_i)^2/n \tag{5}$$

La droite de régression a pour équation:

$$y = b_1 x + b_0 \tag{6}$$

avec

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x} \tag{7}$$

EQUATION DE LA DROITE DE REGRESSION

Exemple des voitures d'occasion

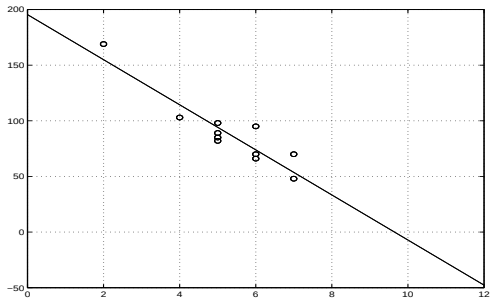

Figure: Abscisse: âge, ordonnée: prix

MISE EN GARDE

Avant de chercher l'équation de la droite de régression, tracer un plot des couples de points (x_i, y_i). Si les points ne semblent pas à peu près alignés, ne pas chercher la droite de régression.

QUANTITES UTILES EN REGRESSION LINEAIRE

- $$SST = \sum_i (y_i - \bar{y})^2 \tag{8}$$
- $$SSR = \sum_i (\hat{y}_i - \bar{y})^2 \tag{9}$$
- $$SSE = \sum_i (y_i - \hat{y}_i)^2 \tag{10}$$

Le **coefficient de détermination** est défini par

$$r^2 = \frac{SSR}{SST} \tag{11}$$

COEFFICIENT DE DETERMINATION

Le coefficient de détermination est toujours entre 0 et 1. Un coefficient de détermination proche de 0 suggère que la droite de régression n'est pas très utile pour effectuer des prédictions. Un coefficient de détermination proche de 1 suggère que la droite de régression est très pertinente.

Exemple des voitures d'occasion:

$$r^2 = 0.853 \tag{12}$$

Interprétation: L'âge de la voiture est très utile pour expliquer le prix de la voiture.

FORMULES UTILES

On a les formules suivantes:

$$SST = S_{yy}, \quad SSR = \frac{S_{xy}^2}{S_{xx}}, \quad SST = SSR + SSE \tag{13}$$

On a aussi:

$$SST = \sum_i y_i^2 - (\sum_i y_i)^2/n, \quad SSR = \frac{(\sum x_i y_i - (\sum x_i)(\sum y_i)/n)^2}{\sum x_i^2 - (\sum x_i)^2/n} \tag{14}$$

$$SSE = SST - SSR \tag{15}$$

3- REGRESSION LINEAIRE MULTIPLE

On suppose à présent que la variable réponse y est *expliquée* par plusieurs variables $x_1, x_2, ...x_p$. Dans l'expérience numéro i, on effectue les mesures $x_{1i}, ...x_{pi}$.

MODELE:

On souhaite *expliquer* la sortie y_i par un modèle dépendant des variables explicatives du type *régression linéaire multiple*

$$y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i} + e_i, i = 1..n \tag{16}$$

où les erreurs e_i suivent une loi normale de moyenne nulle et de variance σ

$$e_i \simeq N(0, \sigma^2) \tag{17}$$

EXEMPLE:Volumes de cerisiers:

Arbre	Diam.	Haut.	Volume	Arbre	Diam.	Haut.	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	16.9	66	64.3
11	11.3	79	24.2	27	17.3	81	55.4
12	11.4	76	21.0	28	17.5	82	55.7
13	11.4	76	21.4	29	17.9	80	58.3
14	11.7	69	21.3	30	18.0	80	51.5
15	12.0	75	19.1	31	18.0	80	51.0
16	12.9	74	22.2	32	20.6	87	77.0

MODELE: PREDICTION DU VOLUME DES ARBRES

But: Prédiction du volume de l'arbre sans l'abattre. On peut mesurer facilement le diamètre de l'arbre et sa hauteur. Si le volume est une variable modélisée en fonction des deux quantitées *diamètre* et *hauteur*, alors on peut estimer de façon simple la valeur économique de la plantation.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1..n \tag{18}$$

PERTINENCE D'UN MODELE DE REGRESSION LINEAIRE

Cas d'une seule covariable.

On trace le graphe de la sortie en fonction de l'unique covariable. On voit sur le graphique si c'est raisonnable.

Cas de plusieurs covariables

On peut tracer la sortie en fonction de chaque covariable prise séparément. Il se peut que chaque graphique montre une relation de type linéaire. Dans ce cas, un modèle linéaire est plausible. Mais il est aussi possible que l'on observe pas de dépendence visuuelle.

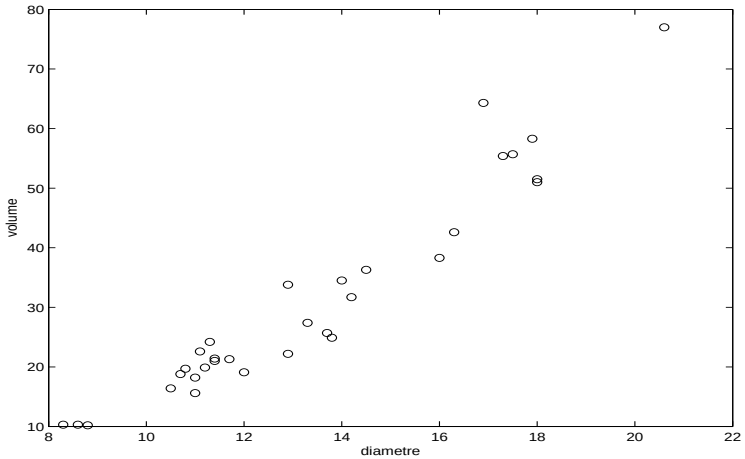

Figure: Distribution des cerisiers diamètre/volume

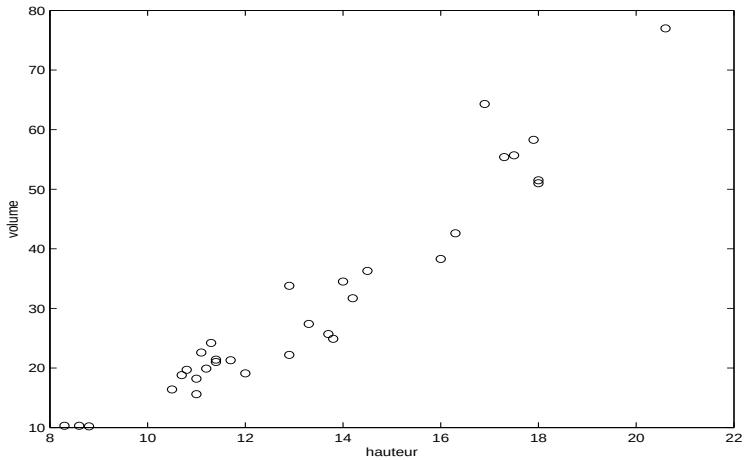

Figure: Distribution des cerisiers hauteur/volume

EXEMPLE DE DONNÉES AVEC 2 COVARIABLES

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_{1i}	2	6	9	14	16	4	10	8	11	3	1	7	13	12	5	15
x_{2i}	17	10	12	25	21	18	14	22	20	11	13	15	24	16	23	19
y_i	29	3	-2	9	-5	25	-1	21	8	14	24	10	10	-3	32	-6

Il n'y a pas de corrélation visuelle entre les variables y et x_1 d'une part, entre y et x_2 d'autre part. Voir la figure. Pourtant on a de façon exacte

$$y = 1 - 3x_1 + 2x_2 \tag{19}$$

REGRESSION LINEAIRE

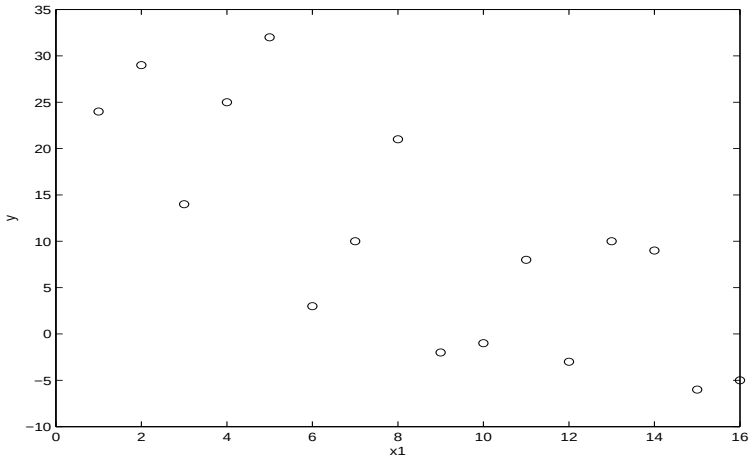

Figure: y en fonction de x1

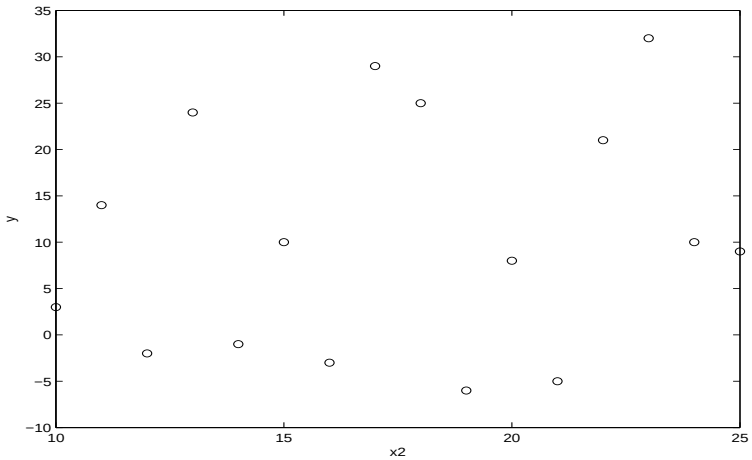

Figure: y en fonction de x2

REGRESSION MOINDRES CARRES AVEC 2 COVARIABLES On cherche un modèle de la forme

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1..n \tag{20}$$

Les 3 coefficients sont évalués par la méthode des moindres carrés. Il s'agit de minimiser la quantité

$$\mathcal{L}(\beta_0, \beta_1, \beta_2) = \sum_i (y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})^2 \tag{21}$$

On obtient après calcul (donné par un logiciel de statistiques) une valeur unique notée $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ qui est

$$\hat{\beta}_0 = -44.8706, \hat{\beta}_0 = 5.1606, \hat{\beta}_2 = 0.0945 \tag{22}$$