Data Analysis

Iw = Sum(m.d²(Point,Centre du gravité du classe auquel il appartient))
Ib = m.Sum(m.Nb_points_classe*d²(Centre de gravité classe, centre de gravité global))
I(G) = Sum(m.d²(point, centre de gravité global)

Matrice de ward :
gamma(Mi,Mj) = (mi*mj / mi+mj) * d²(Mi,Mj)

Modèle Linéaire Simple

Exemple (suite) :

i (ind.)	y_i (revenu)	x_i (âge)	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	52125.0	48.1	193.9	5.0	978.6	25.5
2	50955.9	38.7	-975.3	-4.4	4245.4	18.9
3	53382.9	48.6	1451.7	5.6	8061.1	30.8
4	51286.9	37.5	-644.3	-5.5	3570.3	30.7
5	55243.6	54.7	3312.5	11.6	38434.3	134.6
6	53384.7	40.7	1453.5	-2.4	-3481.4	5.7
7	53488.2	50.1	1557.1	7.1	10982.0	49.7
8	54134.1	45.9	2202.9	2.9	6281.9	8.1
9	52706.4	55.9	775.2	12.9	9975.6	165.6
10	42144.3	25.1	-9786.9	-18.0	176033.4	323.5
11	52665.2	36.9	734.1	-6.1	-4503.3	37.6
12	51656.7	34.5	-274.5	-8.6	2350.7	73.3
Moyenne	*51931.2*	*43.1*	*0*	*0*	*21077.4*	*75.4*
Somme	*623174.0*	*516.8*	*0*	*0*	*252928.4*	*904.3*

cov Var

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{252928.4}{904.3} = \underline{279.7}$$

$$\hat{\beta}_0 = 51931.2 - 279.7 * 43.1 \doteq \underline{39885}$$

12

Beta1_chapeau = cov/var(x)
Beta0_chapeau = y_bar - beta1_chapea*x_bar
Y_chapeau = beta0_chapeau + beta1_chapeau * x
Variance = 1/n * somme (x - xbar)²
Cov = 1/n * somme (x-xbar)(y-ybar)
Coef Corr = cov/racine(variance) * racine(variance)

Calcul SCR :

i (ind.)	y_i (revenu)	x_i (âge)	$\hat{y}_i = 39885 + 279.7 * x_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	52125.0	48.1	53343.0	-1218.0	1483550.6
2	50955.9	38.7	50713.7	242.2	586[illegible]5.3
3	53382.9	48.6	53484.3	-101.4	10274.8
4	51286.9	37.5	50381.1	905.8	820405.6
5	55243.6	54.7	55176.5	67.1	4507.4
6	53384.7	40.7	51261.3	2123.5	4509068.6
7	53488.2	50.1	53903.9	-415.6	172735.6
8	54134.1	45.9	52728.7	1405.4	1975015.2
9	52706.4	55.9	55530.2	-2823.8	7973726.7
10	42144.3	25.1	46900.3	-4756.1	22620189.0
11	52665.2	36.9	50215.3	2450.0	6002285.9
12	51656.7	34.5	49535.7	2121.0	4498484.4
Moyenne	***51931.2***	***43.1***	***51931.2***	***0***	***4177409.1***
Somme	***623174.0***	***516.8***	***623174.0***	***0***	***50128909.0***

$\Sigma \varepsilon$

SCR = (somme des epsilon)²

On note l'expression suivante : $\mathbf{R}^2 = \dfrac{\mathbf{SCE}}{\mathbf{SCT}}$ Avec

$$\mathrm{SCT} = \sum_{t=1}^{n} (y_t - \bar{y})^2 = m_{YY} \qquad SCE = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 = \hat{\beta}_1^2 m_{XX} \qquad \mathrm{SCR} = \sum_{t=1}^{n} \hat{\varepsilon}_t^{\,2} = m_{YY} - \hat{\beta}_1^2 m_{XX}$$

Décomposition de la variance

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

$$SCT = SCR + SCE$$

SCT *: somme des carrés totaux*
SCE *: somme des carrés expliqués par le modèle*
SCR *: somme des carrés résiduels, non expliqués par le modèle*

IV – Validation du modèle : coefficient de détermination R^2

Coefficient de détermination.
Exprime la part de variabilité de Y expliquée par le modèle.
R^2 → 1, le modèle est excellent
R^2 → 0, le modèle ne sert à rien

$$R^2 = \frac{SCE}{SCT}$$

$$R^2 = 1 - \frac{SCR}{SCT}$$

SCE représente la variation expliquée.
SCR représente la variation inexpliquée due aux variables omises dans le modèle.

Si $\mathbf{R^2}$=0,9 ; on dit que 90% de la variation de **X** est expliquée par la variation de Y **.**

Si $\mathbf{R^2}$=0,1 ; la variation de **X** contribue à hauteur de 10% dans l'explication de la variation de **Y.** Par conséquent, la variable explicative ne suffit pas à elle seule à expliquer la variable expliquée **.** On doit dans ce cas introduire d'autres variables dans le modèle sans pour autant rejeter automatiquement la variable X. Ce qu'on appelle modèle linéaire multiple

Modèle linéaire multiple :