

LA REGRESSION LINEAIRE SIMPLE

BUT

TESTER LE TAUX D'INDEPENDANCE ENTRE DEUX VARIABLE QUANTITATIVES

Dans le cas *particulier* où l'on a pu mettre en évidence l'existence d'une *relation linéaire significative* entre deux caractères quantitatifs continus X et Y, on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une des trois équations suivantes :

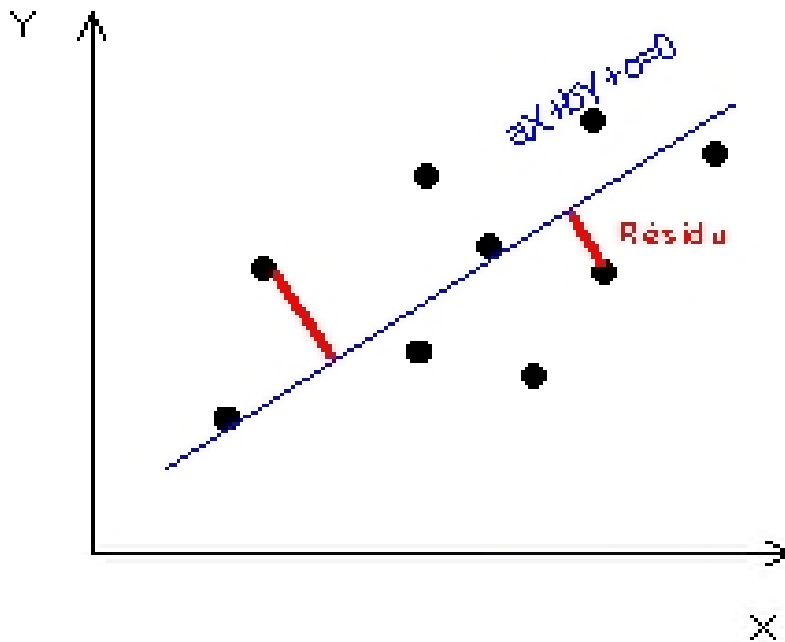
(1) $a.X + b.Y + c = 0$: équation de la **droite moyenne liant les caractères X,Y**

(2) $Y = a.X + b$: **droite de régression de Y en fonction de X**

(3) $X = a.Y + b$: **droite de régression de X en fonction de Y**

Les trois équations proposées ci-dessus correspondent à trois droites différentes, trois résumés différents du nuage de points (X,Y) . La différence entre les trois droites vient du fait que les trois équations proposées correspondent à trois objectifs différents :

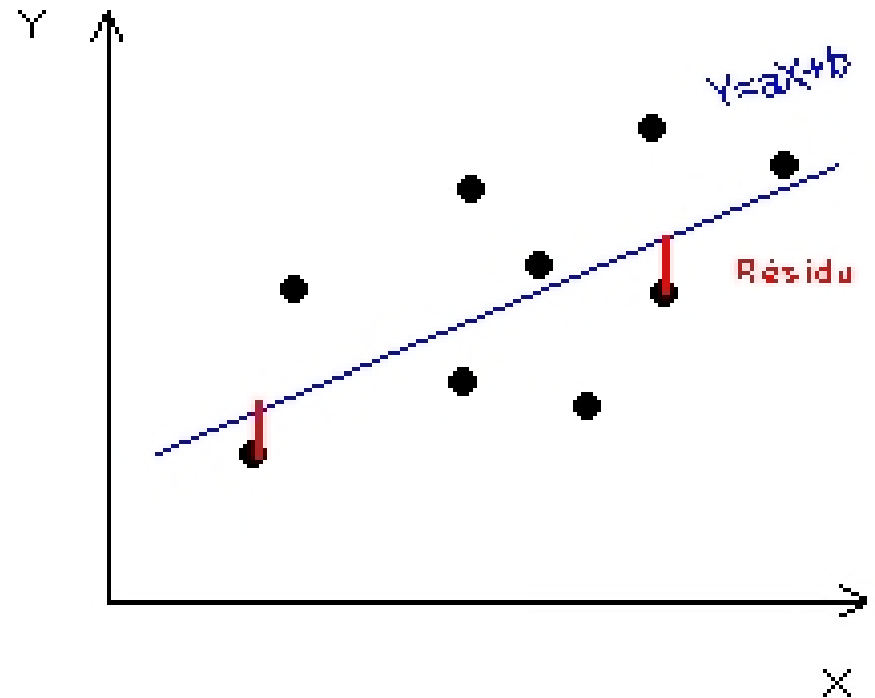
(1) **La droite moyenne** est un résumé de la relation entre X et Y qui n'introduit *aucune hypothèse particulière sur le sens de la dépendance causale* qu'il peut y avoir entre les deux variables. Elle visera donc à tracer la droite qui soit *la plus proche de tous les points*, c'est-à-dire les résidus définis par la perpendiculaire de chaque point à la droite moyenne (plus court chemin).



(1) *Droite exprimant la relation moyenne entre X et Y*

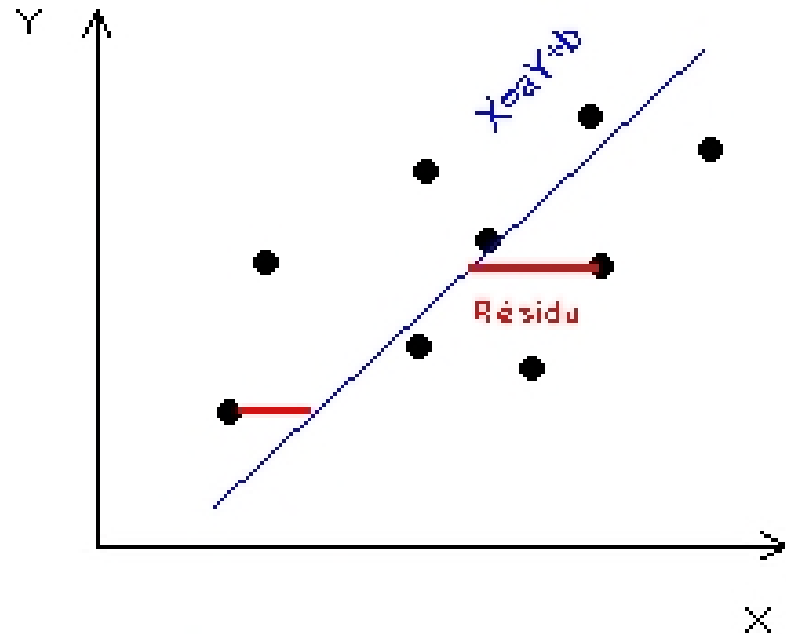
(2) **La droite de régression de Y en fonction de X** introduit l'hypothèse que les valeurs de Y dépendent de celles de X, c'est-à-dire postulent que *la connaissance des valeurs de X permet de prévoir les valeurs de Y*. Il s'agit donc d'un modèle de prévision et l'objectif est de minimiser l'erreur de prévision c'est-à-dire la distance entre les valeurs Y_i observées et les valeurs Y^*_i estimés par la relation $Y^*=aX+b$. Les résidus seront donc la distance à la droite par rapport à l'axe Oy.

(2) *Droite exprimant Y en fonction de X*



(3) **La droite de régression de X en fonction de Y** introduit l'hypothèse inverse que les valeurs de X dépendent de celles de Y, c'est-à-dire postulent que *la connaissance des valeurs de Y permet de prévoir les valeurs de X*. Il s'agit donc cette fois-ci de minimiser l'erreur de prévision sur X c'est-à-dire la distance entre les valeurs X_i observées et les valeurs X^*_i estimés par la relation $X^*=aY+b$. Les résidus seront donc la distance à la droite par rapport à l'axe Ox et non plus par rapport à l'axe Oy comme dans le cas précédent.

(3) *Droite exprimant X en fonction de Y*



7.1 LE CALCUL DE LA DROITE DE REGRESSION $Y=aX+b$

Un exemple pédagogique de régression linéaire

Pour rendre les choses plus claires, nous partirons d'un exemple simple et très classique qui est celui de la relation entre l'altitude (X) et température (Y) à l'intérieur d'une région de taille suffisamment petite pour que l'on puisse négliger les facteurs de variations macroscopiques de la température (distance à la mer, latitude, etc.).

Les données présentées sur la [Figure 2](#) sont imaginaires mais elles pourraient correspondre à la situation d'une vallée alpine de direction nord-sud pour laquelle on a procédé au relevé des températures à midi dans huit stations situées à des altitudes différentes et localisées sur chacun des versants de la vallée.

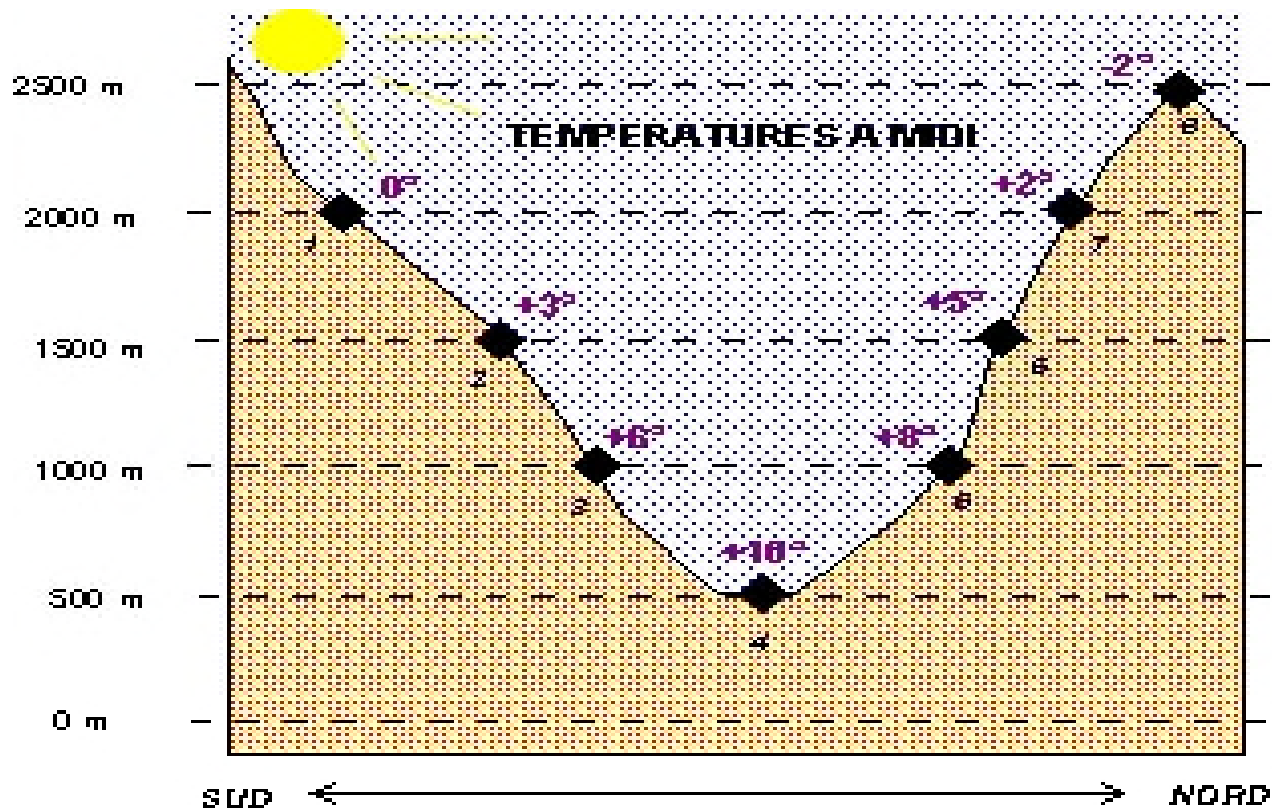


Figure 2 : Température et altitude dans 8 stations (données imaginaires)

Les données relatives à l'altitude et la température des 8 stations peuvent être rassemblées dans un tableau ([tableau 1](#)) à partir duquel on calculera les paramètres caractéristiques de chaque variable (moyenne et écart-type) ainsi que leur covariance.

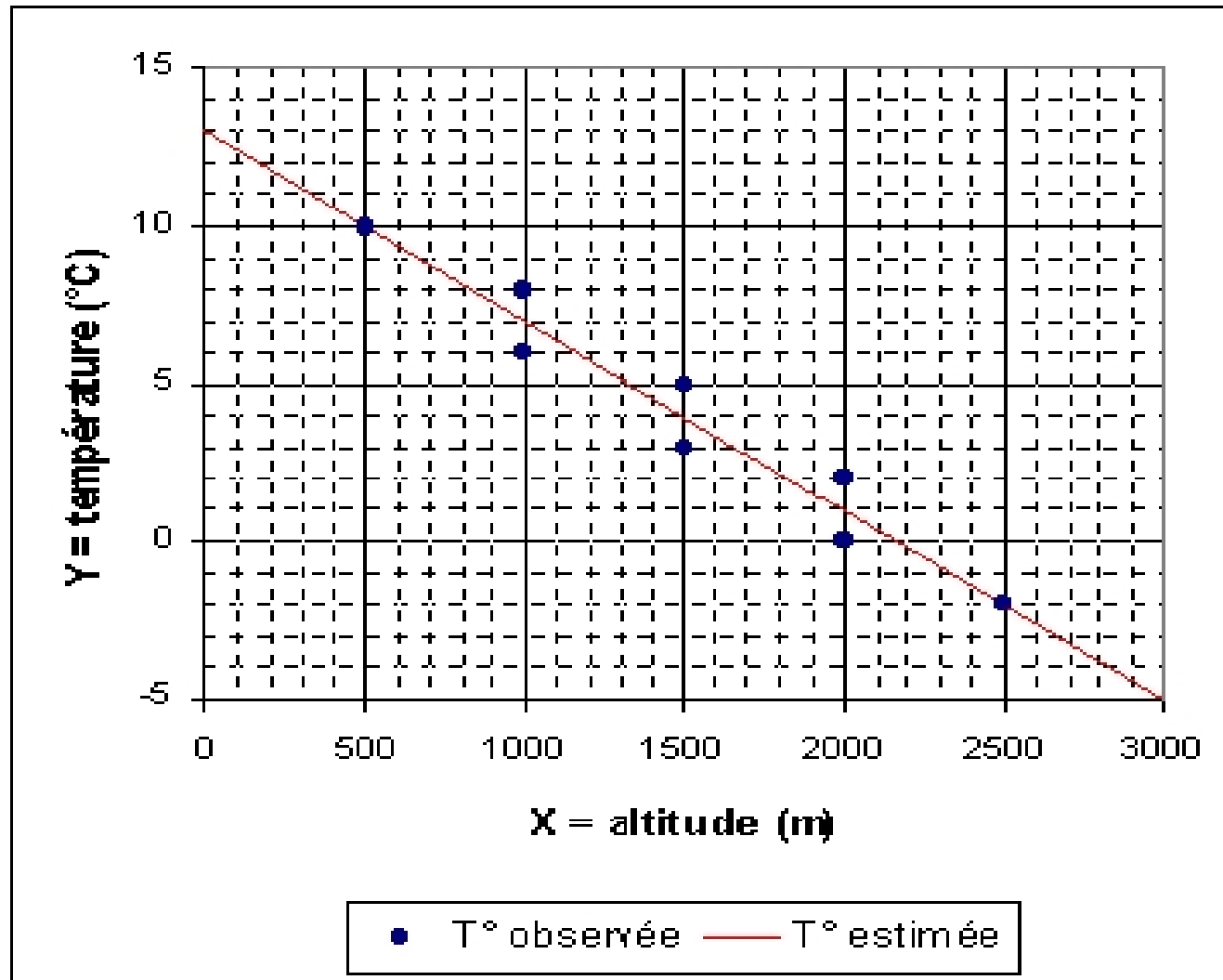
Tableau 1 : Paramètres caractéristiques de la température (Y) et de l'altitude (X) de 8 stations météorologiques (données imaginaires)

i	(Xi) Altitude (m)	(Yi) Température (°)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2000	0	500	-4	-2000
2	1500	3	0	-1	0
3	1000	6	-500	2	-1000
4	500	10	-1000	6	-6000
5	1000	8	-500	4	-2000
6	1500	5	0	1	0
7	2000	2	500	-2	-1000
8	2500	-2	1000	-6	-6000
moyenne	1500	4	0	0	-2250
écart-type	612	3.8	-	-	-

On déduit de la valeur de la covariance (-2250) et de celle des deux écarts-type (612 pour X et 3.8 pour Y) l'existence d'une très forte corrélation linéaire négative entre les deux variables :

$$r(X,Y) = \text{Cov}(X,Y) / [\text{ect}(X) * \text{ect}(Y)] = -2250 / (612 * 3.8) = -0.97.$$

Détermination de la droite de régression par le critère des moindres carrés



Nous avons vu en introduction que lorsque l'on cherche à exprimer Y en fonction de X, on peut affecter à chaque valeur observée Y_i une valeur estimée par la droite de régression $Y^*_i = aX_i + b$. L'erreur d'estimation sur l'individu i est donc égal au résidu ε_i défini par :

$$\varepsilon_i = (y_i - y_i^*) = y_i - (ax_i + b)$$

Comme on souhaite obtenir un ajustement global qui soit optimal pour l'ensemble des stations, il faut définir un critère général définissant la qualité d'ajustement de l'ensemble des valeurs à la droite proposée.

(a) La première solution (ERR1) qui vient à l'esprit est la **minimisation de la somme des résidus** :

$$Err1 = \sum \varepsilon_i$$

Mais ce critère est à l'évidence discutable car des résidus positifs et négatifs peuvent se compenser (températures sous-estimées ou sur-estimées par le modèle) et l'on pourrait obtenir un ajustement optimal ($ERR1=0$) alors même que la droite ne passerait pas par tous les points du nuage.

(b) La seconde solution (ERR2) consiste alors évidemment la **minimisation de la somme des valeurs absolues des résidus** :

$$Err2 = \sum |\varepsilon_i|$$

Il s'agit cette fois-ci d'un bon critère, mais qui a le défaut de ne pas posséder de solution analytique et qui impose une recherche itérative sur toutes les droites du plan.

(c) La troisième solution (ERR3) qui est la plus souvent retenue en statistique est appelée **critère des moindres carrés** et consiste à **minimiser de la somme des carrés des résidus** :

$$Err3 = \sum (\varepsilon_i)^2$$

Comme dans le cas précédent, le critère est correct puisqu'il n'y a pas de compensation entre les résidus positifs et négatifs et la valeur ERR3 ne s'annule que si tous les points du nuage sont alignés le long d'une droite. Mais ce critère possède l'immense avantage d'aboutir à une solution analytique très simple. En effet l'équation de la droite de régression $Y=aX+b$ qui minimise le carré des écarts entre valeurs observées et valeurs estimées est obtenue très simplement à l'aide des deux formules suivantes :

*Les valeurs optimales d'ajustement des paramètres de la droite $Y=aX+b$ **pour le critère des moindres carrés** sont données par les relations :*

$$a = \frac{\text{Cov}(x, y)}{\sigma_x^2}$$

$$b = \bar{y} - a\bar{x}$$

Appliquées aux données du [Tableau 1](#), ces équations permettent d'obtenir les paramètres optimaux d'ajustement de la droite de régression de la température en fonction de l'altitude :

$$a = -2250 / (612*612) = -0.006 \text{ (}^{\circ}\text{C / m)}$$

$$b = 4 - (-0.006 * 1500) = 13 \text{ (}^{\circ}\text{C)}$$

$$\textbf{\textit{Température (}^{\circ}\text{C) = -0.006 * altitude (m) + 13}}$$

$$\begin{array}{lcl} \text{Var}(y) & = & \text{Var}(y^* = ax + b) + \text{Var}(\varepsilon) \\ \text{information totale} & = & \text{information modélisée} + \text{information résiduelle} \end{array}$$

La qualité de l'ajustement correspond donc au rapport entre l'information totale sur Y et l'information effectivement reconstituée à partir de la connaissance procurée par la variable X. Cette qualité d'ajustement varie entre 0% (X n'apporte aucun élément de prévision sur Y) et 100% (la connaissance des valeurs de X permet de prévoir intégralement les valeurs de Y) et dépend de l'intensité de la corrélation entre X et Y. Elle peut se calculer ([Tableau 2](#)) ou se mesurer directement à l'aide du coefficient de détermination, c'est-à-dire du carré du coefficient de corrélation entre X et Y. Si l'on opte pour le calcul, on constate que la variance des températures observées (16.2) est bien égale à la somme de la variance des températures estimées (15.4) et de la variance des résidus (0.9). La qualité d'ajustement est donc égal à $15.4/16.2$ soit 0.95 ce qui correspond également au carré du coefficient de corrélation linéaire des variables X et Y : $(-0.97)^2 = 0.95$.

$$\text{Qualité d'ajustement} = \frac{\text{Var}(y^*)}{\text{Var}(y)} = [r(x, y)^2] = \text{coef. déter}$$

Analyse des résidus d'une régression linéaire

Même si l'importance des résidus d'un modèle de régression est limitée, il est toujours instructif de procéder à leur analyse afin de vérifier : (1) si les résidus ne révèlent pas une mauvaise spécification du modèle utilisé (2) si les résidus ne mettent pas en évidence l'existence d'autres variables explicatives que celle qui a été retenue.

Le premier point sera développé ultérieurement (régression non-linéaire) et l'on va se contenter de développer ici le second à l'aide de l'exemple des températures et de l'altitude.

Tableau 2 : Analyse des résidus de la régression : $\text{Température} = -0.006 \text{ Alt} + 13$

i	(Xi)	(Yi)	$Y^*i=aXi+b$	$Yi-Y^*i$
1	2000	0	1	-1
2	1500	3	4	-1
3	1000	6	7	-1
4	500	10	10	0
5	1000	8	7	1
6	1500	5	4	1
7	2000	2	1	1
8	2500	-2	-2	0
moyenne	1500	4	4	0
variance	428571	16.3	15.4	0.9

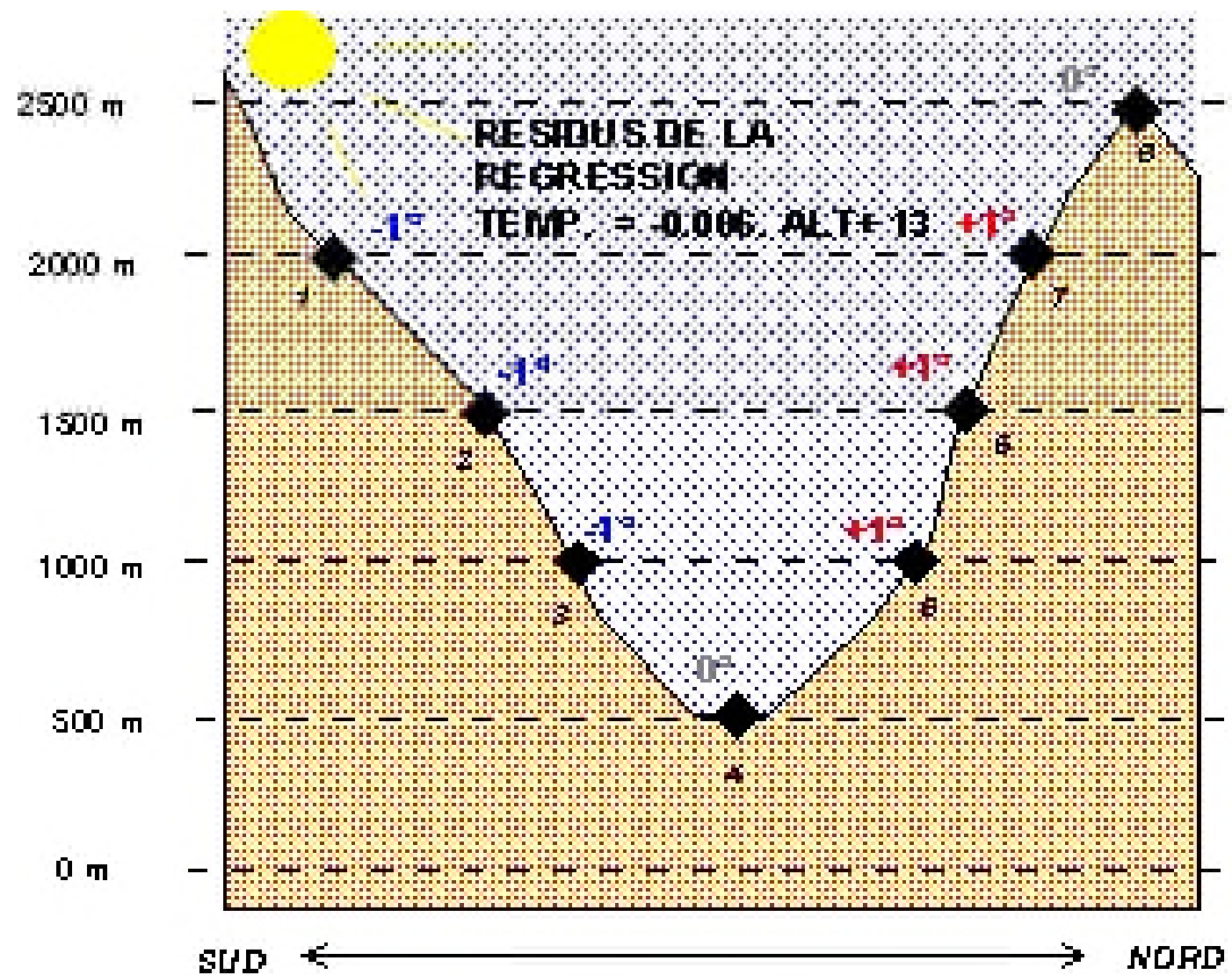
La somme des résidus est nulle (**propriété de la régression linéaire**) mais on constate que trois stations ont des résidus positifs (température réelle supérieure de 1° à ce que laisserait prévoir leur altitude) et trois autres des résidus négatifs (température réelle inférieure de 1° à ce que laisserait prévoir leur altitude).

Comment interpréter ces écarts ?

On peut tout d'abord supposer que ces écarts sont une composante aléatoire liée à l'imprécision des outils de mesure utilisé (précision des thermomètres) mais dans ce cas là la disposition spatiale des résidus devrait être aléatoire. Or, lorsque l'on cartographie les résidus ([Figure 4](#)) on constate que les résidus positifs et négatifs sont loin de se disposer au hasard dans l'espace.

Figure 4 : Configuration spatiale des résidus de la régression :

Température = $-0.006 \text{ Alt.} + 13$



APPLICATIONS PRATIQUES DE LA REGRESSION LINEAIRE

Problème : supposons que l'on dispose de 100 stations météorologiques en Auvergne, pour lesquelles on mesure l'altitude en mètres (X) et la température moyenne (Y) tout au long de l'année.

Est-il réellement utile de retenir chaque jour les 100 températures ?

Réponse : l'observation a montré qu'il y avait une forte corrélation négative (-0.97) entre l'altitude et la température.

La droite de régression $T^{\circ}\text{C} = -0.006.A_m + 10^{\circ}$ permet de résumer l'essentiel de l'information sur la variation spatiale des températures ($-0.9 \times -0.9 = 81\%$), dès lors que l'on connaît l'altitude.

Conclusion : la régression permet de résumer un ensemble volumineux d'informations à l'aide de deux paramètres. Ce résumé est évidemment d'autant plus valable que la corrélation est élevée.

Modéliser

Problème : l'observation répétée tout au long de l'année montre que le coefficient a ne change guère (-0.006) alors que le coefficient b varie selon les saisons (élevé en été et faible en hiver) : que peut-on en déduire ?

Réponse : le coefficient a indique de combien varie la température chaque fois que varie l'altitude. Ainsi, une variation d'altitude de $+100$ m correspond à une diminution de température de $-0.006 \times 100 = -0.6^{\circ}\text{C}$: c'est le **gradient thermique**. Le coefficient b indique la t° correspondant au cas où l'altitude est de 0m : c'est donc la **température moyenne ramenée au niveau de la mer**.