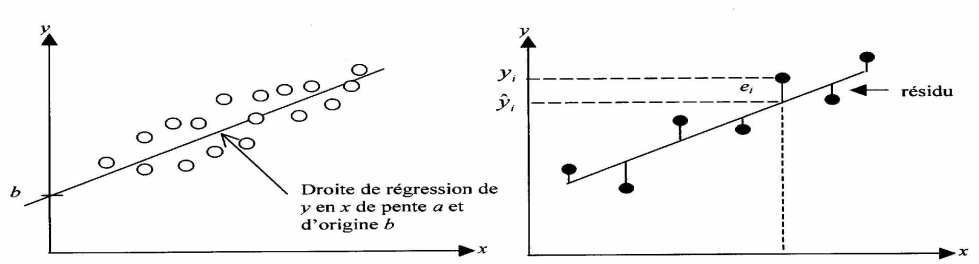
Cours 9 : Régression linéaire

Une situation courante en sciences biologiques est d'avoir à sa disposition deux ensembles de données de taille n, $\{y_1, y_2, \ldots, y_n\}$ et $\{x_1, x_2, \ldots, x_n\}$, obtenus expérimentalement et mesurées sur une même population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x_i et les y_i, par exemple de la forme $y = f(x)$. Lorsque la relation recherchée est affine, c'est-à-dire de la forme $y = ax + b$, on parle de *régression linéaire*. Mais même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs de mesure ou d'une certaine variabilité naturelle, on considère les données $\{y_1, y_2, \ldots, y_n\}$ comme autant des réalisations d'une variable aléatoire Y et parfois aussi les données $\{x_1, x_2, \ldots, x_n\}$ comme autant des réalisations d'une variable aléatoire X et on suppose que la variable Y est, selon le modèle, une fonction affine de la variable X. On dit que la variable Y est la *variable dépendante* ou *variable expliquée* et que la variable X est la *variable explicative* .

1 La droite des moindres carrés

On a vu que des données $\{(x_i, y_i), i = 1, \ldots, n\}$ peuvent être représentées par un nuage de n points dans le plan (x, y). Le *centre de gravité* du nuage se calcule facilement puisqu'il s'agit du point de coordonnées $(\mu(X), \mu(Y)) = (\frac{1}{n}\Sigma_{i=1}^n x_i, \frac{1}{n}\Sigma_{i=1}^n y_i)$. Rechercher une relation affine entre les variables X et Y consiste à rechercher la droite qui s'ajuste le mieux à ce nuage de points. Il n'y a pas une solution unique à ce problème et parmi les droites possibles, le *principe des moindres carrés ordinaire* (MCO) retient celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $\hat{y}_i = ax_i + b$ recherchée. Si ε_i désigne cet écart pour le i-ème point, appelé aussi *résidu*, le principe consiste à choisir les valeurs de a et de b qui minimisent *la somme des carrés des écarts* :

$$\sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} (y_i - (ax_i + b))^2.$$



Un calcul montre que ces valeurs, notées $\hat{a}$ et $\hat{b}$, s'expriment au moyen de la *variance* $\mathrm{Var}\,(X) = \frac{1}{n}\sum_{i=1}^n (x_i - \mu(X))^2$ et de la *covariance* $\mathrm{Cov}\,(X, Y) = \frac{1}{n}\sum_{i=1}^n (x_i - \mu(X))(y_i - \mu(Y))$ des variables X et Y par les formules suivantes[1] :

$$\hat{a} = \frac{\mathrm{Cov}\,(X, Y)}{\mathrm{Var}\,(X)} \;\; et \;\; \hat{b} = \mu(Y) - \hat{a}\mu(X).$$

et la valeur de Y *prédite par la régression* au point x_i est notée $\hat{y}_i$ et vaut $\hat{y}_i = \hat{a}x_i + \hat{b}$.

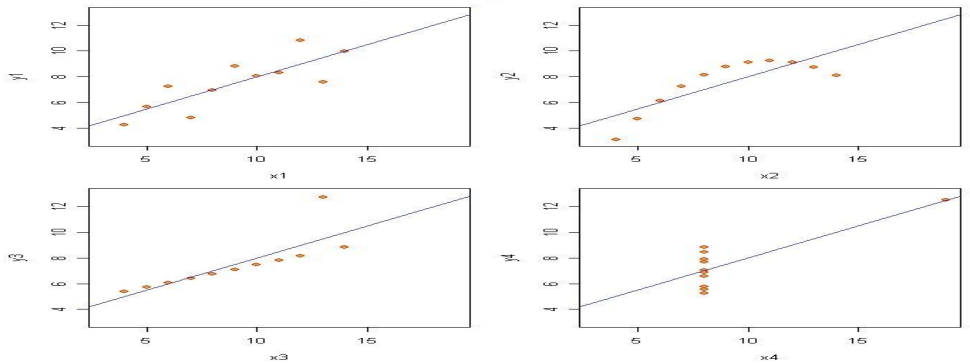
2 Evaluation de la qualité de la régression

Un des plus gros inconvénient de la méthode de régression linéaire est ... qu'elle marche toujours ! C'est en effet un inconvénient car l'utilisateur qui a calculé les deux coefficients $\hat{a}$ et $\hat{b}$ de la droite de régression ne sait pas si celle-ci est un modèle acceptable de ses données ou si ce n'est pas le cas du tout. Pour mesurer la qualité de l'approximation d'un nuage $(x_i, y_i)_{i=1..n}$ par sa droite des moindres carrés, il est donc indispensable de calculer aussi son *coefficient de corrélation linéaire* (qui vaut $\rho(X, Y) = \frac{\mathrm{Cov}\,(X,Y)}{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}$).

[1] A noter que les notations des calculatrices pour les moyenne et variance sont souvent $\overline{x}$ et non $\mu(X)$ et sx^2 et non $\mathrm{Var}\,(x)$.

On a vu en effet que ce nombre, compris entre -1 et $+1$, est une *mesure de la dispersion du nuage*. Il vaudrait $+1$ si les points du nuage étaient exactement alignés sur une droite de pente a positive, et -1 s'ils étaient sur une droite de pente négative. On considére en général que l'approximation d'un nuage par sa droite des moindres carrés est de bonne qualité lorsqu'il est proche de 1 ou de -1 et de médiocre qualité lorsqu'il est proche de 0. Il convient néanmoins de rester prudent comme le montre les exemples suivants :

Exemples : Voici quatre jeux de données choisis de manière à définir la même droite de régression, et avec le même coefficient de corrélation linéaire ρ. De gauche à droite et de haut en bas, le premier jeu de donnée est, au mieux, très bruité mais on peu douter que les donnée soient liées par une relation affine. Le second jeux correspond assez clairement à une relation quadratique : c'est plutôt une courbe $y = ax^2 + bx + c$ qu'il conviendrait d'ajuster. Dans le troisième jeux tous les points sauf un semblent alignés. Il y a visiblement un "point abérrant" dont il faudrait vérifier la provenance (ou la saisie dans le logiciel !) ; la situation est semblable pour le dernier échantillon. Moralité : la régression linéaire donne (presque) toujours une droite, mais il convient de regarder le résultat pour débusquer les situations par trop absurdes.

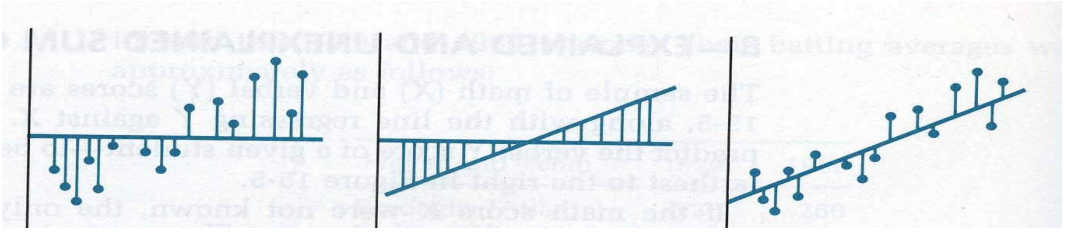


3 Pourcentage de variabilité expliquée par la régression

Parfois les logiciels calculent, avec les coefficients $\hat{a}$ et $\hat{b}$ de la régression linéaire, non pas le coefficient de corrélation linéaire ρ mais son carré, appelé aussi *le R-deux*, $R^2 = \rho^2$ (qui est cette fois compris entre 0 et 1). Le R^2 est important pour la raison suivante : un calcul montre qu'on a la relation

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

qui exprime que la dispersion totale de la série Y (DT) est égale à la dispersion autour de la régression (DA) plus la dispersion due à la régression (DR). Sur la figure ci-dessous où est tracée une droite horizontale qui passe par le centre de gravité du nuage, on voit sur la première figure la dispersion totale DT, sur la seconde la dispersion due à la régression DR (qui est nulle si la pente de la droite des moindres carrés est nulle et grande si cette pente est forte) et sur la troisième la dispersion autour de la droite, ou dispersion résiduelle.



Le lien de cette formule DT=DA + DR avec le R^2 est que l'on a $R^2 = \frac{DR}{DT}$, c'est-à-dire que le R^2 représente *la part de la dispersion totale de Y que l'on peut expliquer par la régression.* Ainsi si l'on obtient une valeur de $R^2 = 0,86$ (et donc $\rho = \mp 0,92$), cela signifie que la modélisation par la droite des moindres carrés explique 86% de la variation totale de Y, ce qui est un très bon résultat.