Aperçu de l'étude des régressions.

Dans un grand nombre de contextes, que ce soit au niveau lycée, supérieur, écriture de thèse ou professionnel ces questions reviennent souvent et il semble bien que ces notions méritent d'être détaillées.

Positionnement du problème.
On dispose d'un ensemble de données organisées de la façon suivante.
Pour un certains nombre de cas, d'individus, d'évènements, on dispose des mêmes observations. Par exemple, pour un certain nombre d'enfants, on aura l'âge, la taille, le poids, le rang dans la famille etc.
Le but est d'établir une formule, que l'on appelle "modèle" et qui résume "au mieux" toutes les informations. Ce modèle pourra être utilisé pour diverses applications, par exemple, la vérification, la prévision, la prise de décision.

On parle souvent de "régression linéaire". L'adjectif "linéaire" signifie que la résolution du problème se fait dans le cadre de l'algèbre linéaire, par opposition à d'autre régressions, par exemple coniques, qui dépassent le cadre de ce papier.

Choix du modèle.
Généralement, on ne peut pas savoir quelle formule adopter pour résoudre au mieux le problème. Mais aussi, on peut imaginer que la formulation du modèle est connu dans sa généralité, et que les données dont on dispose devront permettre de déterminer les paramètres et vérifier le modèle.
Dans le cas général on n'aura pas d'autre choix que d'essayer plusieurs formules et de choisir celle qui donne le meilleur résultat.

Ces généralités étant précisées, il y a lieu de rentrer dans les détails.
Le cas le plus courant est celui de deux variables, c'est à dire que le fonction recherchée est de la forme Y = f(X). Il y a deux variables et il y aura souvent deux paramètres a et b.
La fonction la plus simple et celle qui vient à l'esprit au premier abord est
Y = a + b.X, sa courbe représentative est une droite. Mais on comprend bien qu'il n'y a pas beaucoup de raisons que la variation de Y soit proportionnelle à la variation de X.

On a alors recours à un changement de variable en utilisant les fonctions log et exp.
Compte tenu de cela, avec deux paramètres, ces 5 fonction résolvent la majorité des cas :

```
Ajustement linéaire Y=A + B * X
Ajustement exponentiel Y=A * exp(B * X)
Ajustement logarithmique Y=A + B * ln(X)
Ajustement puissance  Y=A * X^B
Ajustement hyper-logarithmique Y=A + B * ln(X)/X
```

Il y a lieu de remarquer que certaines de ces fonctions ne sont possibles qu'avec des conditions pour X et Y.

Régression avec un polynôme du quatrième degré.

```
Régression polynôme Y=A.X^4 + B.X^3 + C.X^2 + D.X + E
```

On est toujours dans le cadre de la régression linéaire. On a ici 5 paramètres.

Dans certains cas, il y a lieu d'utiliser une fonction exponentielle avec trois paramètres :

```
Ajustement exponentiel Y= A + B * exp(C * X)
```

Le choix de la fonction à adopter ne peut être fait que par comparaison des résidus obtenus pour chaque fonction. Le coefficient R^2 est aussi un bon critère.

Toutes ces formules ne sont valables que s'il y a seulement 2 variables.

Il est courant d'avoir trois variables. La meilleure méthode consiste à combiner deux des quatre premières fonction décrites. On obtient ainsi huit formules possibles, par exemple :
formule : Z = exp(A * X) * Y^B * exp(C)
Plus de détails dépassent le cadre de ce papier. Il est clair qu'un tel traitement n'est envisageable qu'à l'aide d'un programme informatique, mais le principe de base reste toujours le même.

Enfin, reste le cas où le nombre de variables est supérieur à trois.
Il ne paraît pas envisageable et pratiquement sans grand intérêt de combiner plusieurs fonctions.
Dans la littérature, la formule adoptée est de la forme
Y = k + aX1 + bX2 + cX3 + dX4 + eX5 + …
Il est assez peu probable que chaque variable explicative X participe proportionnellement à la variable Y. Il est préférable d'utiliser une formule de la forme :
Y = k . X1^a . X2^b . X3^c . X4^d . X5^e
C'est ce type de formule qui est employé dans les modèles de pluviométrie.
Je tiens à préciser qu'il paraît nécessaire de vérifier la normalité de chacune des variables. En effet, une anomalie dans la liste peut passer inaperçue lors du calcul avec toutes les variables.

Principe et méthode de calcul.
La formule étant fixée, il faut calculer les paramètres.
Quelle que soit la formule, et ceci est vrai dans le cas le plus général, la solution la meilleure, on dit généralement la plus probable, est celle qui minimise la somme des carrés des écarts. Cette somme sera minimale pour les valeurs des paramètres qui annulent les dérivées partielles. Dans les cas favorables, on est ramené à un système linéaire de N équations à N inconnues. Dans les autres cas, il sera nécessaire de recourir à d'autres méthodes calculatoires.

Tout ceci n'est qu'un exposé très rapide et très simplifié de ce qui concerne les régressions linéaires. Il a pour seul but de préciser la problématique et donner quelque détails sur les différents aspects des méthodes à employer.