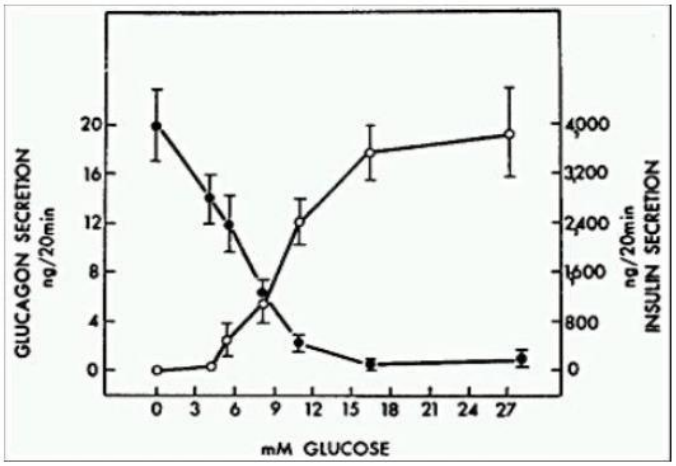
I. LE PROBLEME DE LA NON LINEARITE DES DONNEES.

Il se trouve dans bien des cas que le jeu de données montre un graphique (un "scatter plot") qui de toute évidence ne peut faire l'objet d'une régression linéaire. Ensuite, dans beaucoup de cas, **on sait** que la mesure *y* (variable dépendante) **ne varie pas linéairement** en fonction de *x* (variable explicative). Que faire ?
Ce problème, bien qu'il en relève, est bien plus large que celui de la statistique. Il ne s'agit plus d'établir une seulement une liaison statistique entre deux variables (un coefficient de corrélation, mesurer un effet, s'interroger sur le niveau de signifiance...), mais d'établir un modèle décrivant aussi fidèlement que possible la variation de *y* en fonction de *x* et de l'utiliser à des fins de prédiction. Pour cela, on cherchera une fonction $y = f(x)$, qui ne soit pas du type $Y_i = \beta_1 + \beta_0 X_i + \varepsilon_i$, mais plus complexe (non linéaire). Cette recherche s'inscrit elle même dans une démarche dite de modélisation que l'on rencontre désormais dans tous les domaines de la biologie et des sciences médicales:

http://sites.unice.fr/coquillard/UE10/Strategies%20modelisation.pdf



Taux de sécrétion de glucagon (●) et d'insuline (○) lorsque la concentration de glucose augmente chez le rat (Gerich *et al*, 1974). Chaque point représente la quantité moyenne de sécrétion des deux hormones en fonction de la concentration en glucose.

On voit clairement sur la figure ci-dessus que les concentrations en glucagon ou en insuline ne peuvent être modélisées par une droite.

II. LA REGRESSION NON LINEAIRE.

Premiers exemples voir : http://sites.unice.fr/coquillard/UE7/Regression%20nls.R

Un autre exemple un peu plus compliqué.

On a mesure sur un peuplement d'arbres (le bouleau blanc : *Betula alba*) dans le Massif Central les circonférences des troncs de 21 individus à hauteur de 1.3 m du sol (indice DBH). Dans le même temps, un carottage des arbres a permis d'estimer leurs âges respectifs. Par ailleurs on a constate sur le terrain que les arbres se repartissent en trois catégories : les arbres les plus hauts (dominants), les arbres moyens (co-dominants) et les arbres plus petits, sous le couvert des autres : les dominés.

1. Tracez sur un même graphique les trois courbes représentant la circonférence des troncs en fonction de l'âge. Que constate-t-on et comment interprétez-vous les différences constatées ? Que pensez-vous de l'allure des courbes ? Quel type de fonction peut-on envisager d'ajuster ?

2. On vérifie, en utilisant nls( ) du logiciel R que la croissance en circonférence des troncs peut être modélisé par une exponentielle de saturation de la forme :

$$y(t) = y_{max}(1 - \exp(-rt))^b,$$

où $y(t)$ est la circonférence à l'instant t, y_{max} la valeur maximale théorique que la circonférence peut prendre, r un taux de croissance en circonférence , b un paramètre dit "de forme" de la fonction et t le temps. Les valeurs de y_{max} ont été estimées empiriquement à 86.4 cm, 65.43 cm et 36.00 cm pour chacune des trois catégories d'arbres.

Ages	1	20	40	60	80	100	120
Dominants	1,26	22,29	40,09	56,15	63,49	71,69	81,08
Dominés	1,27	16,02	29,42	31,61	35,61	35,69	35,93
Co-dominants	1,29	22,14	35,69	49,23	56,88	60,43	63,74