Régression linéaire simple

Préparé par:
M. Cherfaoui

Plan du cours

- Position du problème
- Notion de la régression
- Définitions et notations
- Modèle linéaire simple
  - ✓ Estimation des paramètres du modèle
  - ✓ Qualité et Validation du modèle
- Exemples numériques

Position du problème

Dans le cadre de travaux de recherche sur la ***Biomasse*** (*mg*), d'un certain type de plante, en fonction de la concentration de l'Azote NH_4^+ *(µmol)*, nous avons réalisé des expériences dont la biomasse moyenne (*Y*) ainsi que la concertation du l'Azote (*X*) en question sont rangés dans le tableau ci-dessus :

X (concentration *µmol)*	Y (Biomasse *mg*)
0	305
100	378
200	458
400	540
600	565

Y a-t-il un lien entre la biomasse de la plante et la concentration de l'azote?

Notion de la régression

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques, **généralement** pour l'analyse de l'effet d'une ou de plusieurs variables quantitatives sur une variable quantitatives et de l'interprété sous forme d'un modèle mathématique:

$$y = f(x) + \varepsilon,$$

où

- *y*: est une variable quantitative prenant la valeur y_i pour l'individu *i (i = 1,...,n)*, appelée variable à expliquer **(variable expliquée)** ou **variable réponse** (dans **SPSS** elle correspond à la **variable dépendante**).
- *x:* est une variable quantitative prenant la valeur x_i pour le $i^{ième}$ individu, appelées **variables explicatives** ou **prédicteurs** (dans **SPSS** elle représente la **variable indépendante**).
- ε (ε_i pour $i = 1,...,n$): est une variable aléatoire appelée les résidus et elle représente les erreurs du modèle *f* au point x_i (par exemple: elles peuvent être les erreurs de mesure, les effets négligeable d'autres facteurs que *x*,…)

Notion de la régression

Soit $X \in \mathbb{R}^p$

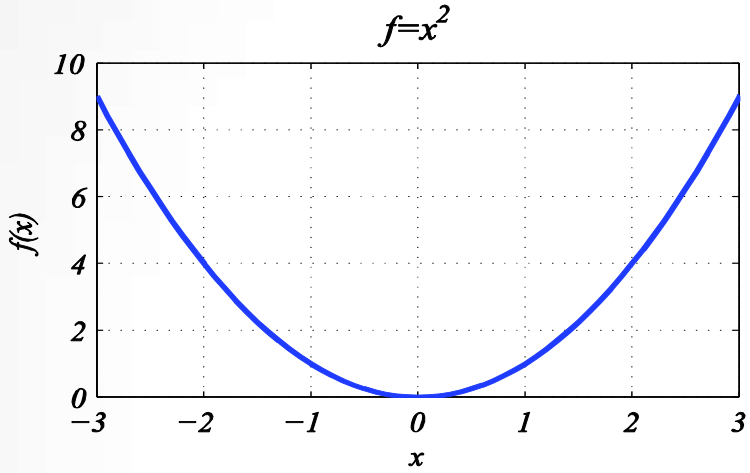
- Si ***p=1*** $(X \in \mathbb{R})$, alors on parlera de la ***régression simple*** en exprimant l'une des deux variables en fonction de l'autre.
- Si $p \geq 2$ $(X=(x_1, x_2, \ldots, x_p))$, alors on parlera de la ***régression multiple***.

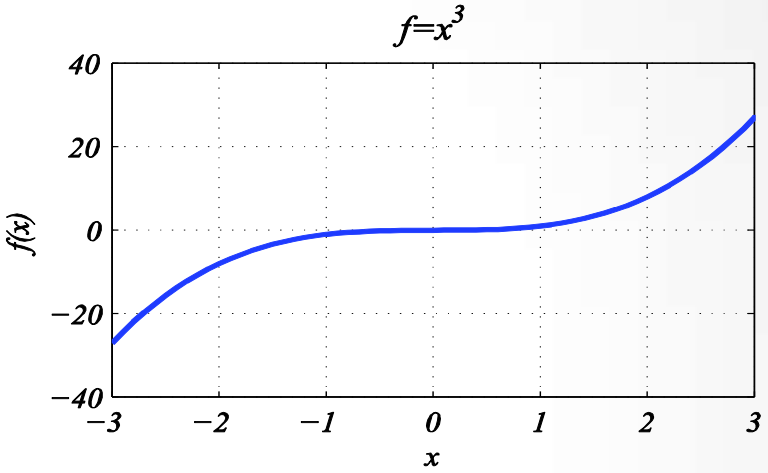
Question: Comment choisir et fixer le modèle de la régression (*f*)?

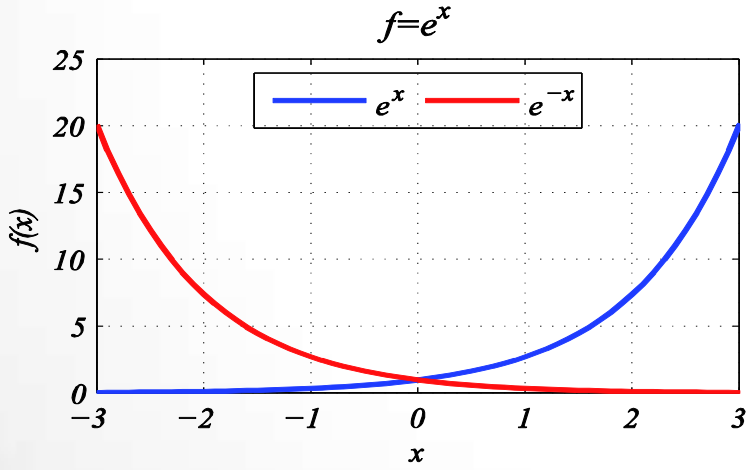
Pour proposer un modèle on doit se référer à l'allure de la distribution du nuage des observations (x_i, y_i).

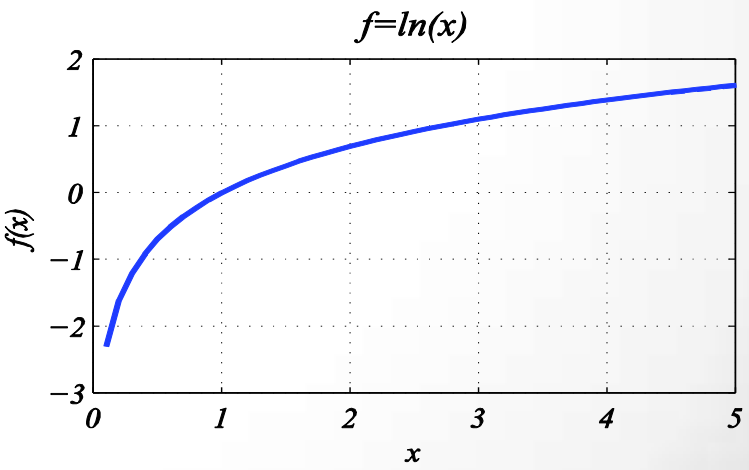
Ceci signifier que, pour proposer un modèle (une fonction paramétrique) proche, dans un certain sens, du comportement du phénomène analysé, la connaissance de l'allure de variation, au moins des fonctions usuelles, est primordiale.

Allure de quelques fonctions usuelles

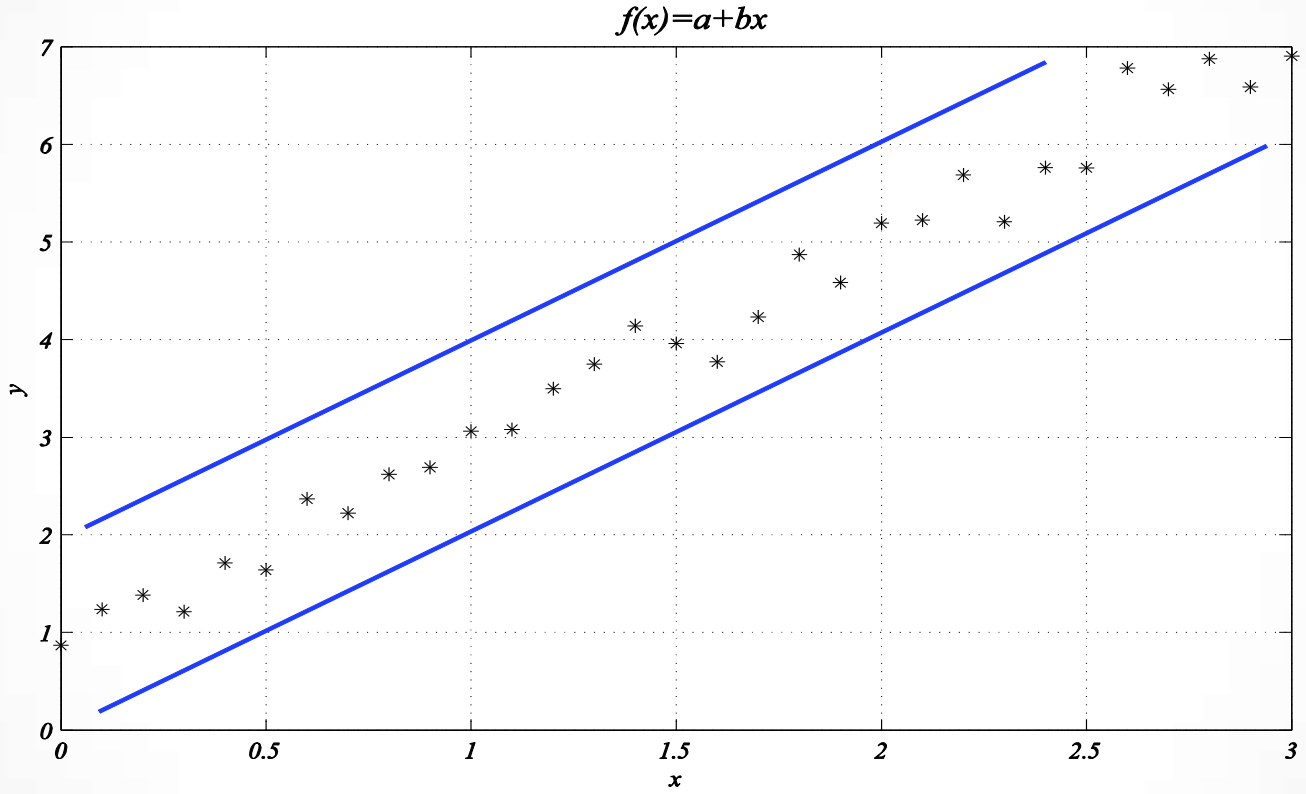








Cas du modèle linéaire



- $f(x)=a + b\,x$ (modèle linéaire simple)
- $f(x)= f(x_1, x_2, \ldots, x_p) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p$ (modèle linéaire multiple).

La linéarité du modèle!

La **linéarité** est liée aux **paramètres du modèle** et non pas aux variables explicatives. En générale, on parle de modèle de régression linéaire si seulement si les paramètres du modèle peuvent être déterminés par **la résolution d'un système d'équations linéaire**.

Exemples:

- $Y_i = a + b\,x_i + c\,x_i^2 + \varepsilon_i$ est un modèle linéaire tandis que la relation entre x et y n'est pas linéaire mais de type polynomial.
- $Y_i = a + b\cos(x_i) + \varepsilon_i$, est un modèle linéaire.
- $Y_i = a\,e^{b\,xi} + \varepsilon_i$, n'est pas un modèle linéaire.
- $Y_i = a\,b + cx_i + \varepsilon_i$, n'est pas un modèle linéaire.

Transformation des modèles non linéaire

Lorsque la dispersion du nuage des points nous suggère un modèle non linéaire souvent on essaye (dans le cas possible) de transformer ce modèle non linéaire vers un modèle linéaire.

Exemples:

- $y=e^{a+bx}$ ⟹ $ln(y)=a+bx$
- $y = ln(a+bx)$ ⟹ $Z=e^{y}=a+bx$
- $y = \frac{a}{b+cx}$ ⟹ $Z= \frac{1}{y}= \frac{a}{b} + \frac{a}{c}\ x$ ⟹ $Z=A+B\ x$
- ....

Une fois que les nouveaux coefficients sont calculés, une transformation inverse doit être réalisée pour revenir au modèle original.

Quelques notions statistiques

- **La moyenne**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **La variance :**

$$\mathrm{Var(x)} = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\bar{X})^2$$

- **L'écart-type:**

$$\sigma_x = \sqrt{\mathrm{Var(x)}}$$

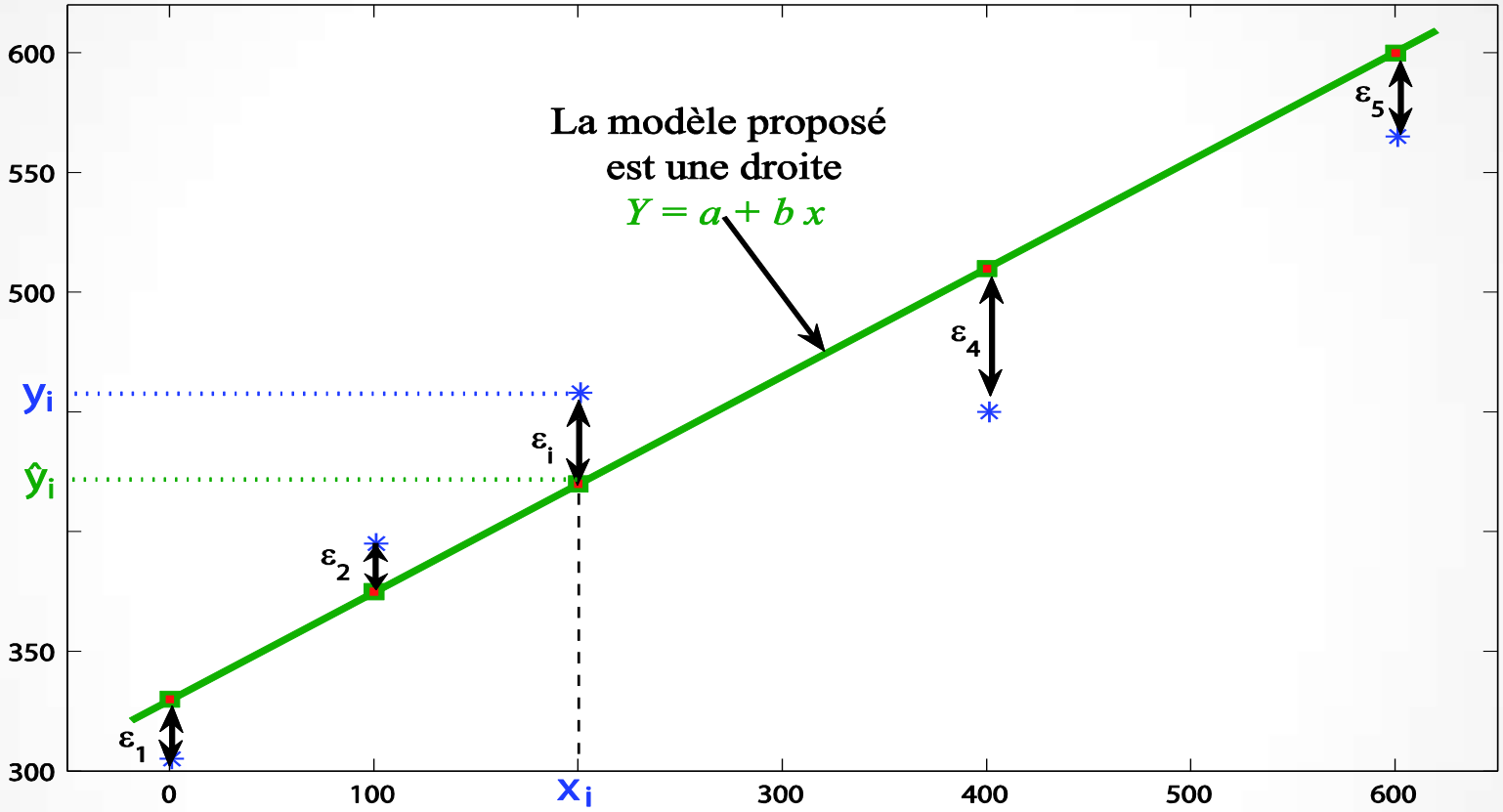
- **La covariance**

$$\mathrm{Cov(x,y)} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - (\overline{XY})$$

- **Coefficient de corrélation linéaire**

$$\rho = r(x,y) = Cor(x,y) = \frac{Cov(x,y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{Cov(x,y)}{\sigma_x \sigma_y}.$$

Estimation des paramètres du modèle linéaire



Afin de déterminer les paramètres d'un modèle il suffit de minimiser les erreurs $\varepsilon_i = Y_i - f(x_i, a, b, c, \ldots)$ **pour** $i=1, 2, \ldots, n$. Pour ce faire, il existe plusieurs techniques, parmi ces dernières on trouve la ***méthode des moindres carrées*** qui consiste à minimiser la somme des erreurs quadratiques.

Estimateurs de a et b via la méthode des moindres carrées

On suppose que la variable X est contrôlée par l'expérimentateur où il réalise n expériences y_1 , y_2 , ... , y_n aux points x_1, x_2, ... , x_n fixés, alors le modèle linéaire en tout point x_i est donné par:

$$y_i = a + b\, x_i \qquad pour \quad i=1, 2, \ldots ,n.$$

Par conspuent, les erreurs du modèle sont définies par:

$$\varepsilon_i = y_i - (a + b\, x_i) \quad pour \;\; i=1, 2, \ldots ,n.$$

Supposons qu'on opte pour la méthode des moindres carrées pour quantifier a et b. Alors les estimateurs des paramètres a et b sont $\hat{a}$ et $\hat{b}$ qui minimise la fonction $Q(a , b)$, définie par:

$$Q(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (a + b\, x_i))^2$$

Estimateurs de a et b via la méthode des moindres carrées

Cela revient à la détermination d'un optimum minimal de la fonction $Q(a\ ,\ b)$, qui consiste à résoudre le système des équations suivant:

$$\begin{cases} \dfrac{\partial Q(a\ ,\ b)}{\partial a} = 0 \\ \dfrac{\partial Q(a\ ,\ b)}{\partial b} = 0 \end{cases} \Longrightarrow \begin{cases} -2\sum_{i=1}^{n}(y_i - a - bx_i) \quad = 0 \\ -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} y_i \ - \sum_{i=1}^{n} a \ - \sum_{i=1}^{n} bx_i = 0 \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} bx_i^2 = 0 \end{cases} \Longrightarrow \begin{cases} a\left(\sum_{i=1}^{n} 1\right) + b\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} y_i \\ a\left(\sum_{i=1}^{n} x_i\right) + b\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} y_i x_i \end{cases}$$

Estimateurs de a et b via la méthode des moindres carrées

La résolution du système précédent, nous fournira la solution suivante:

$$\begin{cases} \hat{b} = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\,\bar{Y}}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{X}^2} \\ \hat{a} = \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - \hat{b}\dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \end{cases} \quad ou\ encore \quad \begin{cases} \hat{b} = \dfrac{Cov(X,Y)}{Var(X)} \\ \hat{a} = \bar{Y} - \hat{b}\bar{X} \end{cases}$$

Ainsi la droite (modèle) de régression sera écrit comme suit:

$$\boldsymbol{\hat{Y} = \hat{a} + \hat{b}x.}$$

Qualité du modèle

Dans ce cours nous allons présenter deux manière du jugé la qualité et l'adéquation du modèle linéaire $y=a + b\ x$ pour l'explication de la variable Y à l'aide de la variable x à savoir:

1. Le coefficient de corrélation et le coefficient de détermination.
2. Le test de Fisher (ANOVA 1).

Qualité du modèle:

Coefficient de corrélation

En probabilité et en statistique, étudier la *corrélation* entre deux ou plusieurs variables aléatoire, c'est étudier l'intensité de la liaison qui peut être existée entre ces variables. Une mesure de cette corrélation dans le cadre linéaire est obtenue par le calcul du *coefficient de corrélation*:

$$\rho = r(x,y) = Cor(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}.$$

- Le coefficient de corrélation est toujours compris entre -1 et +1 ($-1 \leq \rho \leq +1$).

- Le signe de ρ donne le sens de la corrélation entre *Y* et *x*, tel que:

  Si ρ>0 : les deux variables sont proportionnelles,

  Si ρ>0 : les deux variables sont inversement proportionnelles.

Qualité du modèle:

Coefficient de corrélation

- Si $|\rho|$ est près de *1*, alors la corrélation est grande donc le modèle linéaire peut décrire convenablement le phénomène étudié.
- Si $|\rho|$ est près de *0*, alors le modèle linéaire n'est pas adéquat pour la modélisation du problème étudié.

Le coefficient de corrélation nous donne une information sur *l'existence* d'une *relation linéaire* entre les deux variables considérées. A cet effet, un coefficient de *corrélation nul (proche de zéro)* ne signifie pas l'absence de toute relation entre les deux variables mais seulement *l'absence d'une relation linéaire*.

Qualité du modèle:

Coefficient de détermination

Pour mieux juger la qualité d'une régression linéaire, on définit un autre indicateur nommé: *coefficient de* ***détermination***, noté R^2 :

$$R^2 = \rho^2.$$

Ce nombre mesure l'adéquation entre le modèle et les données observées où, plus R^2 est près de +1 plus le modèle est plus adéquat et le contraire est vrai.

Généralement ce coefficient on l'exprime en pourcentage, exemple:

$R^2 = 0{,}86$ on écrit $R^2 = 86\%$.

Qualité du modèle:

Test de validation de Fisher

Une autre technique, plus puissante, que le calcul de coefficient de corrélation, pour mesuré la pertinence et l'adéquation d'un modèle est bien que l'utilisation du test de Ficher ayant le même principe que l'analyse de la variance (ANOVA 1).

On peut démontrer que la variation totale de *Y* se décompose comme suit:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{Y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{Y})^2}_{SCE}$$

Avec: $\hat{y}_i = \hat{a} + \hat{b}\, x_i$, pour $i=1,\ldots,n$ et

SCT : Somme des Carrées de variation de Y ou variation Totale.

SCR : Somme des Carrées de variation des résidus.

SCE : Somme des Carrées de variation de la régression ou variation expliqué par la régression.

Qualité du modèle:

Test de validation de Fisher

Les calculs peuvent être résumé sous la forme d'une table d'ANOVA 1 suivante:

Source de variation	Somme des carrées	*d.d.l*	Carrées moyenne	Ratio
Régression	*SCE*	*1*	*CME=SCE/1*	f_c *=CME/CMR*
Résiduelle	*SCR*	*n-2*	*CMR=SCR/(n-2)*	
Totale	*SCT*	*n-1*		

Décision:

- Si $f_c \leq f(1,n-2,1-\alpha)$ le modèle n'est **pas valide**.
- Si $f_c > f(1,n-2,1-\alpha)$ alors le modèle est **valide**.

La notation $f(1,n-2,1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi de Fisher de degrés de liberté *1* et *n-2* (on l'obtient par la lecture sur la table de Fisher).

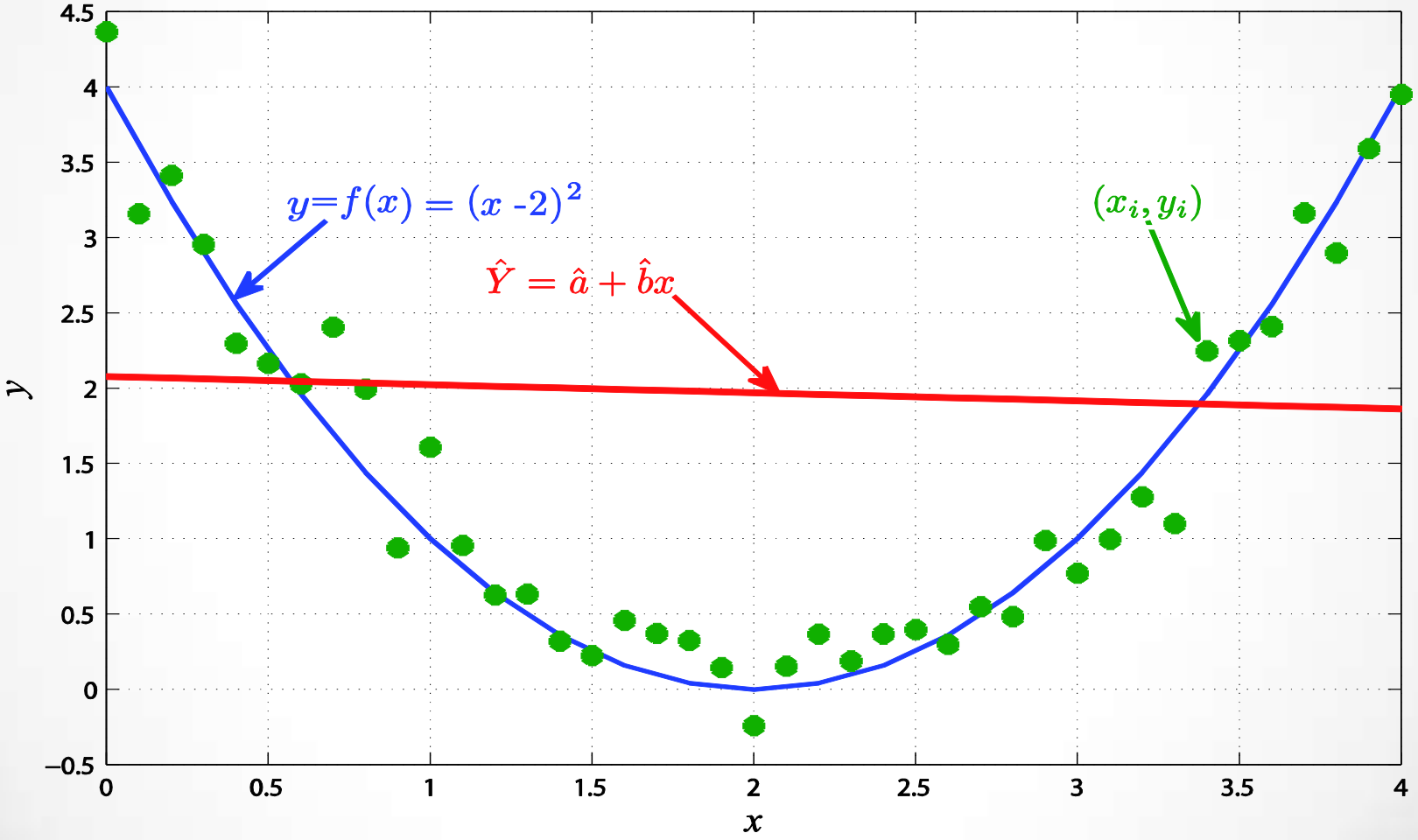
Quelques remarques

1. Si R^2 ou ρ est proche de zéro signifie qu'il n'existe pas un lien linéaire entre les deux variables X et Y mais cela ne signifie guère qu'il n'existe pas un lien entre les deux variables X et Y.
2. Le rejet du modèle (modèle non valide) par le test de Fisher zéro signifie qu'il n'existe pas un lien linéaire entre les deux variables X et Y mais cela ne signifie guère qu'il n'existe pas un lien entre les deux variables X et Y.

La figure suivante est un bon exemple illustratif de ces deux remarques. En effet, on constate que malgré que R^2 et ρ sont très proches de zéro, ce qui signifier que le modèle linéaire $(y = a + b\ x)$ **n'est pas adéquat** pour l'explication de Y à l'aide x et qui est confirmer par le test de Fisher, il reste que le nuage des points indique qu'il existe un **lien quadratique** entre Y et x, où $Y = (x-2)^2$.

ρ=-0.056, R^2 =0.003 (pas d'un lien linéaire)

f_c=0.1236 < $f(n_1,n_2,1-0.05)$= 4.0913 (modèle non valide)



Exemples d'application

- **Voir la série TP N° 3 (avec solution).**

Fin