# La régression linéaire

## Compréhension et pratique

# Sommaire

# I. Interpretation Géométrique

Dans cette partie, nous allons aborder la régression linéaire simple du point de vue géométrique. Pour celà nous allons commencer par transfomer nos échantillons en vecteur.

Vu que nous avons n échantillons de la forme $(x_i, y_i)$, soient les vecteur suivant :

$$Y = [y_1, .., y_n] et\ X = [x_1, .., x_n]$$

Or, pour tout i, nous avons supposé qu'il existe a,b deux réels vérifiant :

$$y_i = ax_i + b + \epsilon_i$$

Cette relation devient donc :

$$Y = aX + b\mathbb{I} + \mathrm{E} \text{ avec } \begin{cases} \mathbb{I} = [1, .., 1] \\ \mathrm{E} = [\epsilon_1, \dots, \epsilon_n] \end{cases}$$

Or, $\mathrm{E}.\mathbb{I} = \sum_{i=1}^{n} \epsilon_i = 0$ et $\mathrm{X}.\mathbb{I} = 0$ car, par hypothèse, X et $\mathbb{I}$ sont idépendants.

De plus les vecteur $\mathbb{I}\ et\ X$ ne sont pas colinéaires, donc il forment une base de $R^n$ de dimension 2. Soit $\tilde{Y} = aX + b\mathbb{I}$. $\tilde{Y}$ est donc le projeté orthogonal de $Y$ sur le plan formé par la base $(X, \mathbb{I})$ (vu que le vecteur $\mathrm{E}$ est orthogonal aux vecteurs $\mathbb{I}$ et $X$).

D'où, par définition du projeté orthogonal, $\tilde{Y}$ minimise la distance entre Y et le plan $(X, \mathbb{I})$.

Ce qui veut dire :

$$d(Y, \tilde{Y}) = \min_{Z \in P} d(Y, Z)\ \ avec\ P\ le\ plan\ formé\ par\ (X, \mathbb{I})$$

Donc, on rejoint le résultat obtenu par la méthode des moindres carrés :

$$\|Y - \tilde{Y}\|^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \min_{c,d \in R^2} \sum_{i=1}^{n} (y_i - (cx_i + d))^2$$

## II. Cas d'erreurs gaussienne

Dans le cours précédent, des hypothèses, dites faibles, ont été faites sur les erreurs $\epsilon$.

$$(\mathcal{H})\begin{cases} E(\epsilon_i) = 0 \ \forall \ i \in \{1,..,n\} \\ Var(\epsilon_i) = \sigma^2 \ \forall \ i \in \{1,..,n\} \\ Cov(\epsilon_i, \epsilon_j) = 0 \ si \ i \neq j \end{cases}$$

Afin d'obtenir de meilleurs résultats de prédiction, nous allons effectuer des hypothèses plus lourdes sur ces erreurs (qu'il faudrait vérifier ultérieurement bien sur) :

$$(\mathcal{H}')\begin{cases} \epsilon_i \sim N(0, \sigma^2) \forall i \in \{1,..,n\} \\ Cov(\epsilon_i, \epsilon_j) = 0 \ si \ i \neq j \end{cases}$$

Cette hypothèse faite, la loi des $y_i$ devient donc :

$$y_i \sim N(ax_i + b, \sigma^2)$$

Et comme les $\epsilon_i$ sont mutuellement indépendants donc les $y_i$ le sont aussi.

Connaissant donc la loi régissant les $\epsilon_i$ et les $y_i$ nous pouvons donc predire plus sûrement en imposant des intervalles de confiance et en calculant de mainère plus précise nos estimateurs.

(Pour plus de détails sur la suite de la procédure, voire le cours (1) mis en référence à la fin de ce document)

## III. L'utilisation du log

Malheureusement, les cas exposés depuis le début de ce cours sont des cas peut fréquents lors de l'analyse de données réelles et s'approchent d'une situation idéale plus qu'une situation réelle (les relations linéaires entre la variables à prédire et la ou les variable(s) explicatives sont assez rares).

C'est pour celà que les transformations semi-logarithmiques ou bi-logarithmiques sont utilisés, afin de ramener des modèles potentiellement non-lineaires, ou bien qui ne vérifient pas les hypothèses faites sur les résidus, au modèle linéaire que nous venons d'étudier.

Les deux méthodes consistent à mener l'études linéaire en remplaçant les couple (X,Y) par :

$$(X, \ln(Y)) \quad (transformation\ semi - \text{logarithmique})$$

$$(\ln(X), \ln(Y)) \quad (transformation\ bi - \text{logarithmique})$$

Ces deux transformations permettent de linéairiser (grace aux propriété du log) des relations quadratiques, exponentielles, ... Et donc d'obtenir des meilleurs résultats de prédiction après application de la régression linéaire pour ces types de relations sans pour autant erroner les résultats si la relation entre X et Y est bel et bien linéraire.

## Références

(1) http://www.lsta.upmc.fr/guyader/files/teaching/Regression.pdf

(2) http://grasland.script.univ-paris-diderot.fr/STAT98/stat98_7/stat98_7.htm