## 1. Introduction

Plusieurs techniques issues de la statistique et de la probabilité ont permis d'accroître les connaissances sur l'analyse de données, la suppression de données aberrantes ou gérer les données manquantes pour choisir une représentation pertinente d'un phénomène. Une fois les données bien préparées, se pose la question de comment tirer des informations efficaces sur des données en grande quantité qui nécessiterait des procédures trop gourmandes en ressources informatiques et des connaissances peu développées. C'est à ce niveau qu'intervient l'apprentissage automatique qui permet de rendre un programme capable d'apprendre à partir d'exemple de données sans être programmé. Une fois l'algorithme implémenté, la machine peut apprendre et prédire des phénomènes précis et s'enrichir au fur et à mesure qu'il reçoit de nouvelles données. L'un des concepts de base de l'apprentissage automatique est la régression linéaire.

## 2. La régression linéaire

L'objectif de la régression linéaire est d'exprimer une variable de sortie y en fonction de la variable d'entrée x de manière linéaire, c'est à dire $y=ax+b$. Ce modèle a donc deux paramètres A et B, dont il faut trouver les valeurs optimales durant la phase d'apprentissage. Prédire la valeur d'une maison en fonction de sa superficie, sa localisation, la possibilité de parking ou non, prédire le nombre d'utilisateurs et utilisatrices d'un service en ligne à un moment donné sont deux exemples d'utilisation du modèle de régression linéaire.
Plusieurs techniques existent pour estimer ces paramètres, les plus répandues étant la méthode des moindres carrés, la méthode des déviations et la méthode du maximum de vraisemblance.
La régression linéaire est un algorithme d'apprentissage supervisé, on dispose alors de N couples entrée-sortie constituant l'ensemble de données $D=\{xi,\}i\in[1,N]$. L'objectif est de trouver une fonction dite de prédiction ou une fonction coût qui décrit la relation entre ***X*** et ***Y*** c'est-à-dire qu'à partir de valeurs connues de ***X***, on arrive à donner une prédiction des valeurs de ***Y***. La fonction recherchée est de la forme :

***Y= f(X)*** avec f***(X)*** une fonction linéaire

À partir d'un échantillon de population qui représente nos données, on répartit les données en deux groupes, les données d'entraînement et les données de test. La première catégorie de données servira pendant la phase d'apprentissage du modèle alors que le second sera utilisé pour évaluer la qualité de prédiction du modèle. Le but n'est donc pas de construire une fonction qui prédira avec une précision optimale les valeurs des variables cibles mais une fonction qui se généralisera au mieux pour prédire des valeurs de données qui n'ont pas encore été observées. Avant de débuter une étude de régression simple, il faut d'abord tracer les observations. (***$X_i$, $Y_i$), i=1,…, p***

## 2.1. Représentation graphique

Le but est de savoir si le modèle linéaire est oui ou non pertinent pour l'étude de notre phénomène. Le graphique est au départ un nuage de points et on relève la tendance qu'a la forme de ce nuage de points.

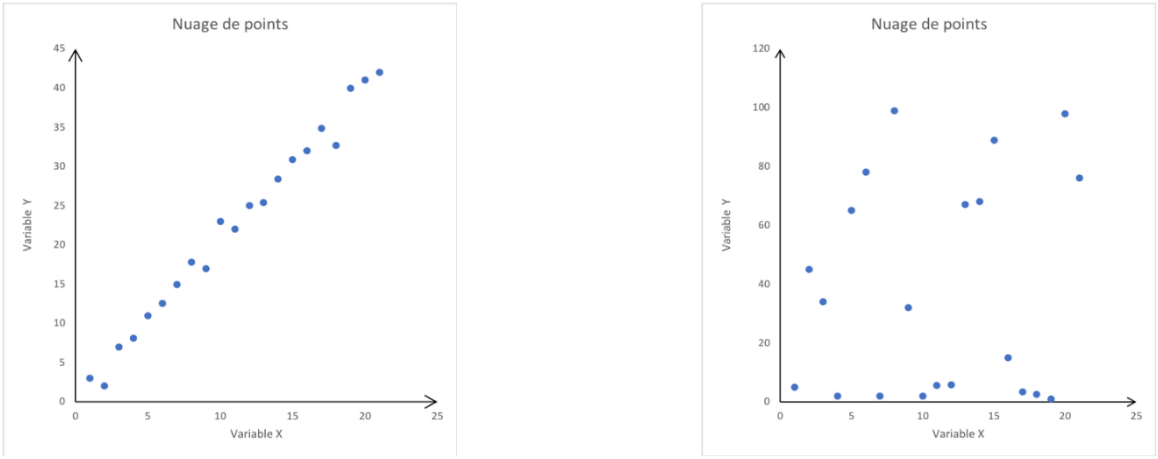


Figure 1. Représentation d'un nuage de points

Au vu de ces deux graphiques, il semble approprié d'utiliser le modèle linéaire pour la première image et pas pour la deuxième qui ne laisse transparaitre aucune tendance connue.Dans la suite nous expliquerons la modélisation et l'estimation des paramètres de la fonction de prédiction pour pouvoir tracer cette droite.

## 2.2. Modélisation

### A. Modèle de la régression linéaire

| **Modélisation** | **Nature de la régression** |
|---|---|
| Une seule variable explicative *X* | Régression simple |
| Plusieurs variables explicatives $X_j$ *(j=1,…, q)* | Régression multiple |

Le modèle de régression linéaire analyse les relations entre la variable dépendante ou variable cible ***Y*** et l'ensemble des variables indépendantes ou explicatives ***X***. Cette relation est exprimée comme une équation qui prédit les valeurs de la variable cible comme une combinaison linéaire de paramètres.

### B. Un modèle de régression linéaire simple est de la forme :

$Y = f(X) + \varepsilon$ où $f(X) = aX+b$

Donc $Y = aX+b+ \varepsilon$

Avec :

- ***Y***, la variable cible, aléatoire dépendante
- ***a*** et ***b***, les coefficients (pente et ordonnée à l'origine) à estimer
- ***X***, la variable explicative, indépendante
- ***ε***, une variable aléatoire qui représente l'erreur

**C. Un modèle de régression linéaire multiple est de la forme :**

$$Y=ax_1+bx_2+cx_3+\cdots+K+\varepsilon$$

Avec :

- ***Y***, la variable cible, aléatoire dépendante
- ***a***,…,***K*** les coefficients (pente et ordonnée à l'origine) à estimer
- $X=(x_1,\ldots,x_p)$, la variable explicative, indépendante
- ***ε***, une variable aléatoire qui représente l'erreur

**D. Sous forme matricielle, le modèle de régression linéaire simple est de la forme :**

$$Y=AX+\varepsilon$$

Où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ & \vdots \\ 1 & x_p \end{pmatrix}, A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

Avec :

- ***Y***, un vecteur à expliquer de taille ***p*** **x 1**,
- ***X***, la matrice explicative de taille ***p*** **x 2**,
- ***ε***, le vecteur d'erreurs de taille ***p*** **x 1**

***ε*** est appelé résidus c'est l'erreur commise, c'est-à-dire l'écart entre la valeur $Y_i$ observée et la valeur $a_iX_i+b$ donnée par la relation linéaire. En effet, même si une relation linéaire est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour ce faire, on tient compte dans le modèle mathématique des erreurs observées.

## 2.3. Principe de fonctionnement

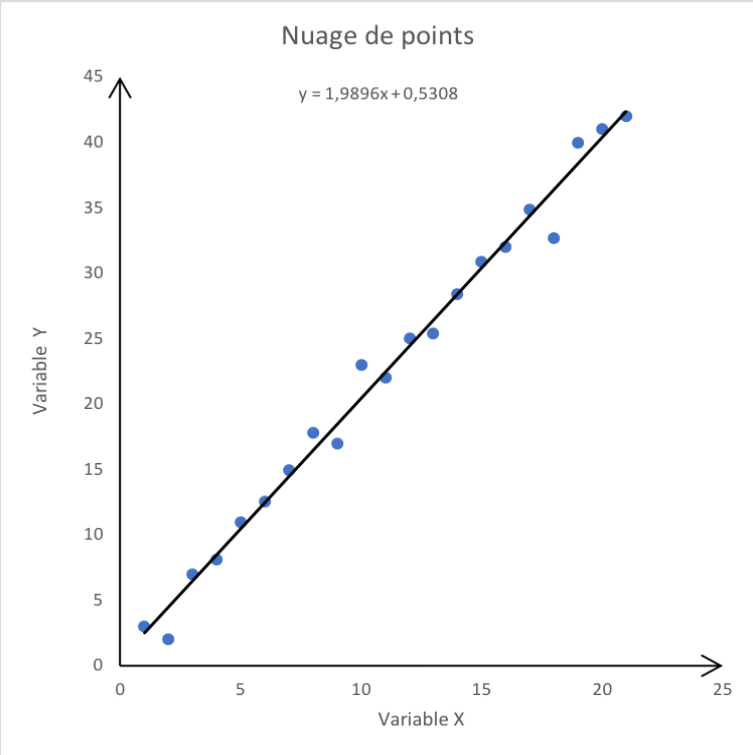


Figure 2. La droite de régression linéaire

Sur ce graphique, la droite de régression linéaire ou la droite des moindres carrés de ***Y*** en ***X*** représente la droite d'ajustement linéaire, celle qui résume le mieux la structure du nuage de points pendant la phase d'apprentissage. Elle rend minimale la somme des carrés des erreurs d'ajustement.
C'est en confrontant l'équation calculée par l'algorithme de régression linéaire aux nouvelles données de la réalité (X) que les prédictions (Y) seront réalisées par l'algorithme d'intelligence artificielle.
Le terme ***r(X,Y)*** représente le coefficient de corrélation de Bravais-Pearson. Ce coefficient mesure l'intensité de la relation linéaire entre ***Y*** et ***X*** . Ce coefficient est calculé à partir des écarts types $\sigma x$ et $\sigma y$ des variables et à partir de la covariance entre les variables d'entrée et de sortie. Voici sa formule :

$$R = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{p}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{p}(Y_i - \bar{Y})^2}}$$

Pour simplifier :

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \,.\, \sigma_y}$$

Le coefficient de corrélation est un nombre toujours compris entre -1 et 1.

- Si ***R*** est proche de 1 : il y a une forte liaison linéaire entre les variables et les valeurs prises par ***Y*** ont tendance à croître quand les valeurs de ***X*** augmentent.
- Si ***R*** est proche de 0 : il n'y a pas de liaison linéaire
- Si ***R*** est proche de -1 : il y a une forte liaison linéaire et les valeurs prises par ***Y*** ont tendance à décroître quand les valeurs de ***X*** augmentent.

### 2.4. Estimation des coefficients de la droite par la méthode des moindres carrés

Dans la figure suivante, deux modèles de régression linéaire : le premier modèle présente des écarts importants entre les valeurs prédites et attendues tandis que le second minimise les carrés de ces écarts

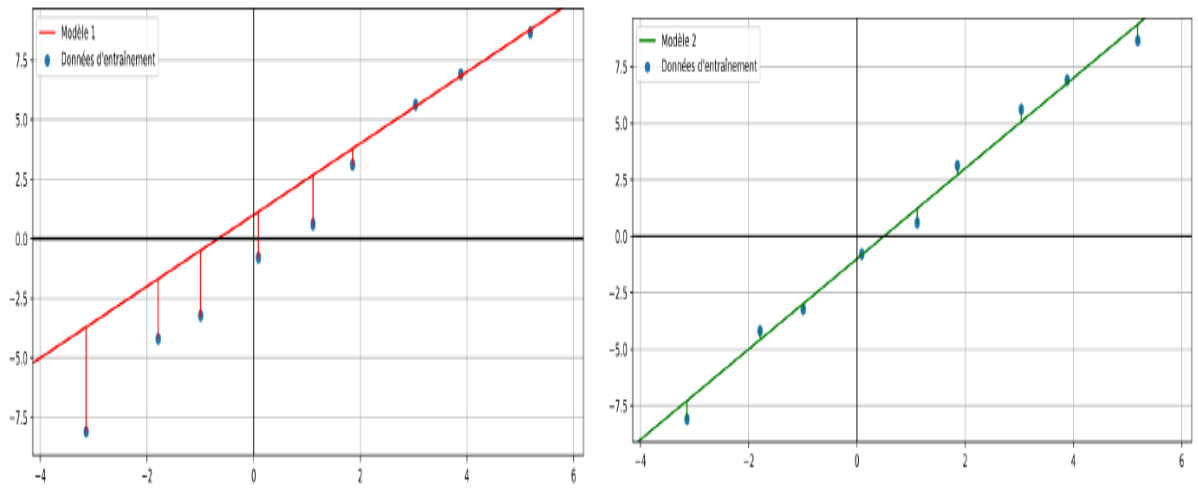


Figure 3. Le choix de la droite de régression linéaire.

La régression linéaire est relativement simple d'un point de vue mathématique. Ce qui fait que ce type d'algorithme entre pleinement dans le cadre d'apprentissage automatique, est le fait que la machine soit capable d'ajuster les paramètres ***a*** et ***b*** à partir d'exemples fournis par l'utilisateur. Dans cette partie, nous expliquons comment ces paramètres sont ajustés afin d'estimer la variable de sortie ***Y***.

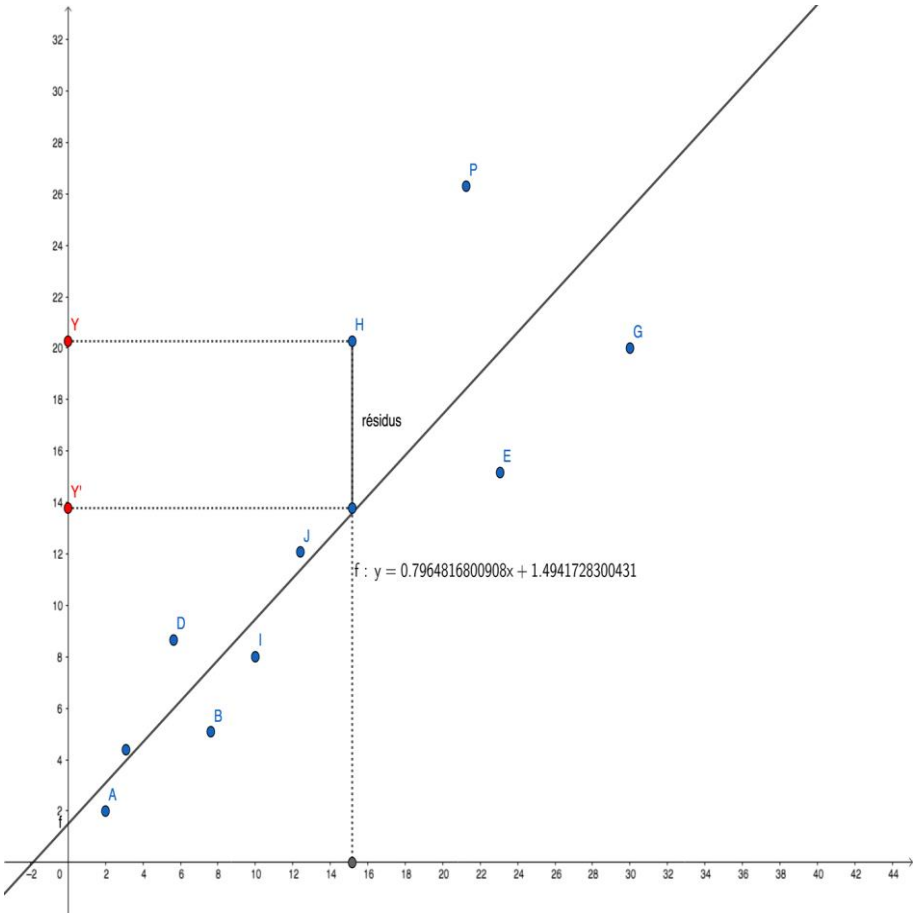


Figure 4. Principe des moindres carrés pour la régression linéaire.

Le principe des moindres carrés ordinaires consiste à choisir les valeurs de **a** et **b** qui minimisent les erreurs de prédiction ou les résidus sur un jeu de données d'apprentissage :

$$\varepsilon = \sum_{i=0}^{P} \left(Y_i - (aX_i + b)\right)^2$$

Minimiser cette expression revient à résoudre un problème d'optimisation, voici la forme des estimateurs notés **a** et **b** qui sont égaux à :

$$a = \frac{Cov(X,Y)}{Var(X)}$$

$$b = \bar{y} - a\bar{x}$$

Où **Cov(X, Y)** est la covariance entre les $\boldsymbol{X_i}$ et les $\boldsymbol{Y_i}$ et **Var(X)** est la variance des $\mathbf{X_i}$. L'expression de **b** indique que la droite de régression linéaire passe par le centre de gravité du nuage de points **($\bar{\mathbf{X}}$, $\bar{\mathbf{Y}}$)**.

## 3. Exemple pratique de régression linéaire

Pour rendre les choses plus claires, nous partirons d'un exemple simple et très classique qui est celui de la relation entre l'altitude (X) et température (Y) à l'intérieur d'une région de taille suffisamment petite pour que l'on puisse négliger autres facteurs de variations de la température (distance à la mer, latitude, etc.). Les données sont présentées dans le tableau suivant :

| i | (Xi) | (Yi) |
|---|---|---|
| **1** | 2000 | 0 |
| **2** | 1500 | 3 |
| **3** | 1000 | 6 |
| **4** | 500 | 10 |
| **5** | 1000 | 8 |
| **6** | 1500 | 5 |
| **7** | 2000 | 2 |
| **8** | 2500 | -2 |

À partir du tableau on calculera les paramètres caractéristiques de chaque variable.

| i | (Xi) | (Yi) | $Xi^2$ | $Yi^2$ | Xi.Yi |
|---|---|---|---|---|---|
| **1** | 2000 | 0 | 4000000 | 0 | 0 |
| **2** | 1500 | 3 | 2250000 | 9 | 4500 |
| **3** | 1000 | 6 | 1000000 | 36 | 6000 |
| **4** | 500 | 10 | 250000 | 100 | 5000 |
| **5** | 1000 | 8 | 1000000 | 64 | 8000 |
| **6** | 1500 | 5 | 2250000 | 25 | 7500 |
| **7** | 2000 | 2 | 4000000 | 4 | 4000 |
| **8** | 2500 | -2 | 6250000 | 4 | -5000 |
| **moyenne** | 1500 | 4 | 2625000 | 30,25 | 3750 |

On déduit de la valeur de la covariance (-2250) et de celle des deux écarts-type (pour X et pour Y) l'existence d'une très forte corrélation linéaire négative entre les deux variables :

$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x . \sigma_y}$ = -2250 / (612 * 3.8) = -0.97.

La forme du nuage de point croisant les valeurs de X et de Y est par ailleurs parfaitement linéaire ce qui justifie la recherche d'un ajustement à l'aide d'une droite.
Il reste à déterminer le **sens de la relation**, c'est-à-dire l'hypothèse faite sur la variable explicative (indépendante) et la variable à expliquer (dépendante). Dans l'exemple choisi, il paraît assez naturel de supposer que la température (Y) dépend de l'altitude (X) et non pas l'inverse, de sorte que l'on va chercher à la température Y en fonction de l'altitude X.

- **Détermination de la droite de régression par le critère des moindres carrés**

Dans l'exemple qui est proposé, on devine facilement le tracé de la droite de régression qui donnera le meilleur ajustement des températures en fonction de l'altitude mais il faut se munir d'un critère objectif pour démontrer que la solution proposée est bien la solution optimale, critère que l'on pourra ensuite appliquer à des nuages de points plus complexe où la détermination de la droite de régression optimale est moins évident.

En appliquant la méthode la plus souvent retenue en statistique **critère des moindres carrés** qui consiste à minimiser de la somme des carrés des résidus, on aura les valeurs optimales d'ajustement des paramètres de la droite Y= aX+b:

$$a = \frac{Cov(X,Y)}{Var(X)}$$

$$b = \bar{y} - a\bar{x}$$

Appliquées aux données, ces équations permettent d'obtenir les paramètres optimaux d'ajustement de la droite de régression de la température en fonction de l'altitude :
**a** = -0.006 (°C / m)
**b** = 13 (°C)
On en déduit que l'équation générale donnant la température en fonction de l'altitude dans l'exemple étudié est la suivante :

***Température (°C) = -0.006 * altitude (m) + 13***

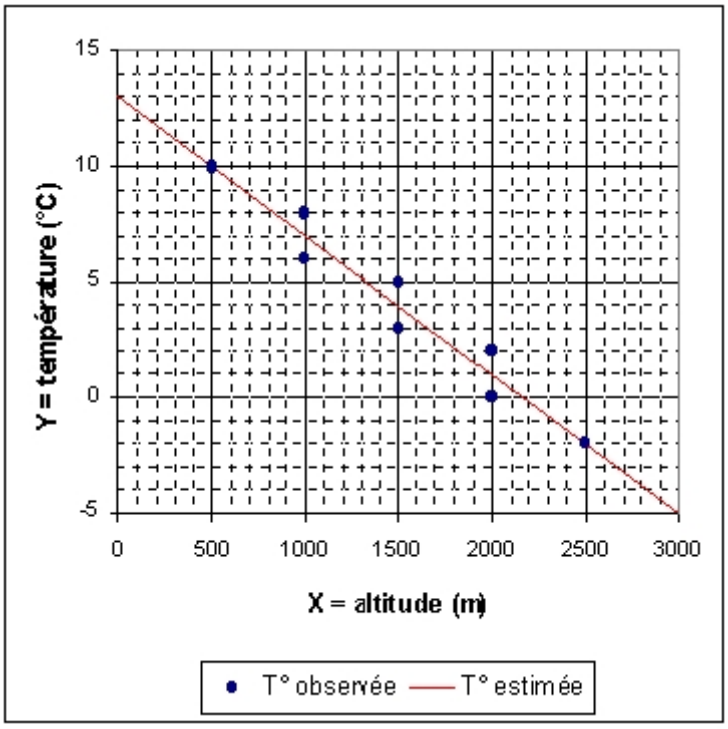


Figure 5. Droite de régression exprimant la température en fonction de l'altitude.

- **Signification des paramètres de la droite de régression**

Le paramètre **a** de la droite de régression indique de combien varie en moyenne la valeur de **Y** lorsque celle de **X** augmente d'une unité. Dans notre exemple, la valeur de **a** est égale à -0.006 et indique que la température diminue en moyenne de 6 ° C chaque fois que l'altitude augmente de 1000 mètres. D'un point de vue géométrique, la valeur de **a** correspond à la pente de la droite de régression par rapport à l'axe **Ox**.

Le paramètre **b** de la droite de régression correspond quant à lui à la valeur théorique de **Y** lorsque la valeur de **X** est égale à 0. Dans notre exemple, il s'agit donc de la température estimé pour une altitude nulle. D'un point de vue géométrique, la valeur de **b** correspond à la coordonnée verticale de l'intersection entre la droite de régression Y=aX+b et l'axe **Oy**.

L'interprétation empirique des paramètres a et b dépend évidemment de la nature des variables X et Y mises en relation, mais les principes définis précédemment demeurent valable en tout état de cause : *a* est le taux de variation de Y en fonction de X et *b* est la valeur de Y pour X =0. Ainsi, dans le cas d'une régression temporelle du type *Y(t)=a.t+b*, le paramètre *a* correspond au taux moyen de croissance (variation de Y par unité de temps) et *b* à la valeur de Y au temps t=0.

## 4. Conclusion

L'avantage de l'algorithme de régression linéaire est sa simplicité d'interprétation et sa facilité de calcul. Par contre, le data scientist veillera à bien vérifier qu'il existe une relation linéaire entre les paramètres d'entrée et celle de sortie. Le modèle présente quelques inconvénients comme le fait que l'algorithme est très sensible aux valeurs aberrantes (outliers) des données d'apprentissage d'où la nécessité de bien préparer ses données dès le départ. Il existe des méthodes dites de régularisation pour pallier à ce problème. Les méthodes de régularisation permettent de pénaliser les valeurs trop grandes des coefficients $\boldsymbol{a_i}$ et $\boldsymbol{b}$.