****

**Regression Lineaire**

Realize par: NOUBHANI Mohammed

AIT DAOUD Salaheddine

**Encadré par:** Pr. Ghazdali

Sommaire:

* + - **Motivation**
    - **Définition**
    - **Objectif**
    - **Formulation de la problématique**
    - **Etude de quelque exemples**
    - **Test de l’ajustement**
    - **Conclusion**

I Introduction

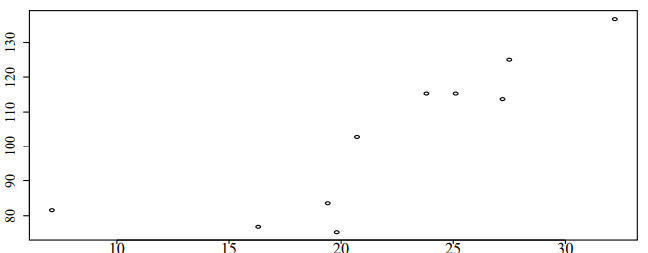
Le but de la régression simple (resp. multiple) est d'établir un lien entre une variable dépendante Y et une variable indépendante X (resp. plusieurs variables X1, ..., Xq). pour pouvoir ensuite faire des prévisions sur Y lorsque X est mesurée.

II Représentation graphique

Le but de la régression simple est de chercher une fonction f telle que

yi ≈ f(xi)

Pour définir ≈, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. Ainsi une étude de régression simple débute toujours par un tracé des observations  
 (xi, yi), i = 1, ..., n. Cette première représentation permet de savoir si le modèle linéaire est pertinent.



En prenant le graph ci-dessus , la modélisation par une droite de la relation entre X et Y semble correspondre à une bonne approximation de la liaison.

Un modèle de régression linéaire simple est de la forme : Y = β0 + β1X + ε avec :

• Y est la variable d´dépendante (une v.a.).

• β0 et β1 sont les coefficients (ordonnée `a l’origine et pente).

• X est la variable indépendante (variable explicative).

• ε est une erreur aléatoire.

L’espérance de Y pour chaque X est le point sur la droite d’équation E(Y |X) = β0 + β1X.

On suppose que :

• Pour chaque valeur de X, E(ε) = 0 et V(ε) = σ

• ε ∼ N(0, σ2).

• Les erreurs ε sont indépendantes (non corrélées).

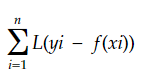
On cherche à :

• Estimer les paramètres β1, β2 et σ.

• Vérifier si le modèle est adéquat.

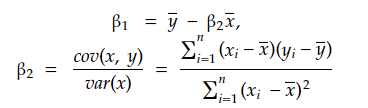
III - Moindres Carrés Ordinaires

1. méthode direct :

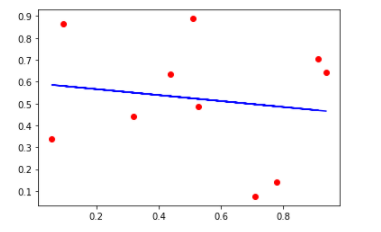
Les points (xi, yi) étant donnés, le but est maintenant de trouver une fonction affine f telle quela quantité :

soit minimale. Pour pouvoir déterminer f, encore faut-il préciser la fonction de coût L.

On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO) β1 et β2 les valeurs minimisant la quantité :

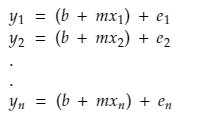
Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (xi, yi) du nuage à la droite ajustée y = β1 + β2x .Les estimateurs des MCO ont pour expressions :

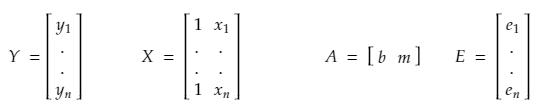
Code en python :



1. **Méthode linéaire :**

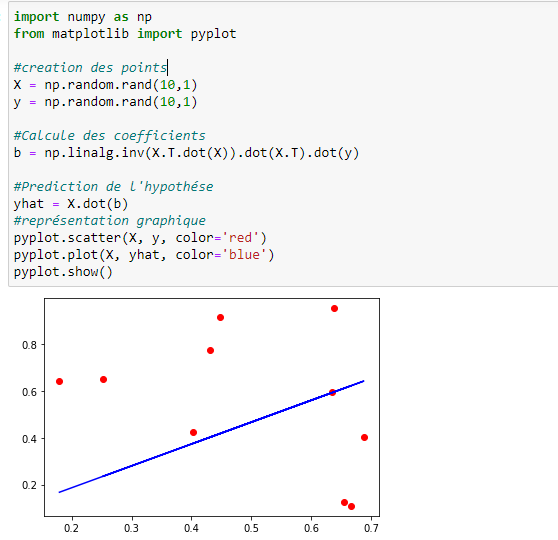
la résolution de l’équation y =mx + b revient à la résolution du système suivant



on posant :

notre problème devient la résolution de l’équation : Y = XA + E

avec :

L’implémentation en python :

Avantages

+ Pas de taux d'apprentissage

+Pas d'itérations

+Fonctionne très bien lorsque le nombre de fonctionnalités est inférieur.

Les inconvénients :

-coûteux en calcul lorsque le jeu de données est volumineux.

-Lent lorsque le nombre de fonctionnalités est supérieur

-Le temps de fonctionnement est O (n³)

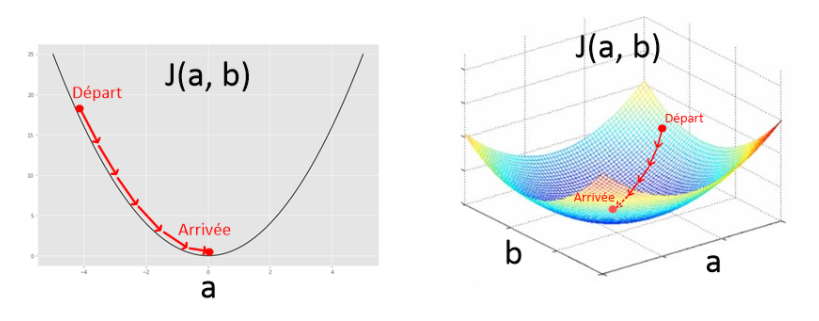
-Parfois, votre X transpose X est non inversible, c'est-à-dire une matrice singulière sans inverse.

**IV - Descente de gradient**

la descente de gradient convergera lentement notre hypothèse vers un minimum global, où le coût serait le plus bas. Ce faisant, nous devons définir manuellement la valeur alpha, et la pente de l’hypothèse change par rapport à la valeur de notre alpha. Si la valeur de alpha est grande, cela prendra de grands pas. Sinon, dans le cas du petit alpha, notre hypothèse convergerait lentement et par petits pas .

Nous partons d’un point initial aléatoire puis nous mesurons la valeur de la pente en ce point En calculant la **dérivée** de la fonction.

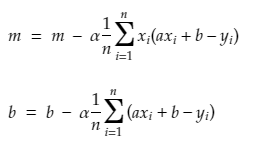
On progresse ensuite d’une certaine distance dans la direction de la pente qui descend On appelle cette distance ***Learning Rate***, que l’on pourrait traduire par *vitesse d’apprentissage.*

En répétant ces deux étapes en boucle, l’algorithme de Gradient Descente est donc un algorithme **itératif**. Pour l’illustrer sur un graphique, je vais prendre l’exemple de la fonction coût que l’on a développé pour une régression linéaire. L’algorithme permet de trouver la valeur idéale pour les paramètres m et b.

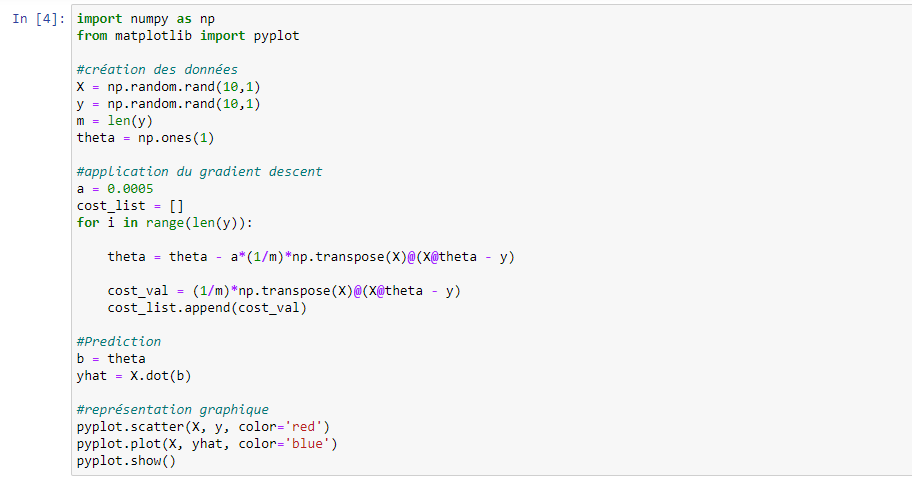
On peut écrire la fonction coût comme :

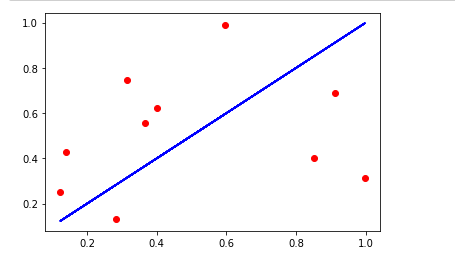
L’équation du gradient s’écrit comme :



Ce qui nous donne :

et maintenant tout ce que nous avons à faire est de répéter jusqu'à la convergence

algorithme en python :



Les avantages importants de la descente de gradient sont

+Le temps de fonctionnement est de O (kn²)

+Fonctionne bien avec plus de fonctionnalités

Les inconvénients importants de Gradient Descente sont

-Besoin de choisir un taux d'apprentissage α

-Nécessite de nombreuses itérations pour converger

-Peut être bloqué dans les minima locaux

-Si le taux d'apprentissage α n'est pas correct, il se peut qu'il ne converge pas.

V - Valeurs ajustées, résidus et somme des carrés des résidus

Une fois les coefficients de la droite estimés, on calcule

- pour chaque individu :

\_ s'appelle la valeur ajustée ou prédite de *Y* par le modèle.

\_ s'appelle le résidu de l'observation *i*. C'est l'écart entre la valeur de *Y* observée sur l'individu *n*°*i* et la valeur prédite. Le résidu *ei* est une approximation du terme d'erreur *εi*.

\_ la somme des carrés des résidus est . Elle mesure la distance de la droite de régression aux

points du nuage de points qui est minimale au sens des moindres carrés.

\_ La statistique est un estimateur sans biais de *σ*².

qualité de l'ajustement

la variation totale des observations *yi* autour de leur moyenne :

peut être décomposée en deux parties :

où représente la variance expliquée par la régression (mesure la variation des

valeurs ajustées autour de la moyenne ) et représente la variance résiduelle ou non

expliquée (partie de la variation totale qui n'est pas expliquée par le modèle de régression).

Afin d'avoir une idée globale de la qualité de l'ajustement linéaire, on définit le coefficient de

détermination qui est le carré du coefficient de corrélation *R :*

Il mesure la part de la variation totale de *Y* expliquée par le modèle de régression sur *X*.

Cas particuliers :

- si *R*2 = 0, le modèle n'explique rien, les variables X et Y ne sont pas corrélées linéairement.

- si *R*2 = 1, les points sont alignés sur la droite, la relation linéaire explique toute la variation.

- une valeur de *R*2 proche de 1 est nécessaire pour avoir un ajustement raisonnable mais en aucun cas suffisant.

Tests

a ) Test global de significativité de la régression

Il paraît raisonnable de tester la significativité globale du modèle, c'est à dire tester si tous les coefficients sont supposés nuls, excepté la constante.

Cela correspond dans le cas de la régression linéaire simple à *H*0 : *b*1 = 0 contre *H*1 : *b*1 *̸*= 0

La statistique du test : statistique *F* de Fisher

On utilise la statistique, notée *F* définie par la formule :



Loi de *F* sous *H*0

La statistique *F* suit la loi de Fisher à (1*, n −* 2) ddl.

Région de rejet de *H*0

Sous *H*0, on s'attend à observer une valeur de *F* proche de 0. Plus la valeur de *F* est grande et plus elle est en faveur de *H*1.

La région de rejet est située à l'extrémité droite du domaine.

Décision

Règle basée sur la p-valeur : si *αobs ≤ α*, on rejette *H*0 au risque d'erreur *α*.



Les valeurs observées de *F* sont données ainsi que la p-valeur.

Régression linéaire multiple :

Dans de nombreuses applications, il y’a plus d'un facteur qui influence la réponse. Les Modèles de régression multiples décrivent ainsi comment une seule variable de réponse Y dépend linéairement d'un certain nombre de variables prédictives.

ie : Le prix de vente d'une maison peut dépendre de l'opportunité de l'emplacement, le nombre de chambres, le nombre de salles de bain, l'année de construction , la superficie en pieds carrés du terrain et un certain nombre d'autres facteurs.

Le modèle de régression linéaire multiple est une généralisation du modèle de régression simple. Lorsque les variables explicatives sont en nombre quelconque. Nous supposons donc que les données

collectées suivent le modèle suivant :



Généralement , Un modèle de régression linéaire multiple est défini par une équation de la forme :



Avec :

• Y est un vecteur aléatoire de dimension n,

• X est une matrice de taille n × p connue, appelée matrice du plan d’expérience,

• β est le vecteur de dimension p des paramètres inconnus du modèle,

• est le vecteur de dimension n des erreurs.

Même S'il est possible d'estimer les paramètres de modèles linéaires plus complexes avec des méthodes similaires à celles que nous avons les calculs deviennent rapidement très compliqués.

Estimateurs des Moindres Carrés Ordinaires :



et la matrice PX de projection orthogonale sur M(X) s’écrit :

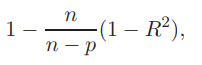


Les résidus sont définis par:

Le coefficient de détermination R² est défini par :



Le coefficient de détermination ajusté R² a est défini par :



Conclusion

La régression linéaire simple et la régression linéaire multiple (MLR) sont des méthodes simples et usuelles pour établir un modèle linéaire entre une variable réponse quantitative, et une ou plusieurs variables explicatives. Cependant, il faut étre trés vigilant à toujours observer le modéle et les résidus

pour trouver d’éventuels écarts aux hypothèses faites : hétéroscédasticité, non normalité des résidus, dépendance des résidus... Dans le cas de la régression multiple, s’il y a de fortes corrélations entre les prédicteurs et/ou si le nombre de prédicteurs est supérieur au nombre d’échantillons, le modèle n’est pas adapté. On peut alors utiliser une méthode de sélection de variables pour réduire le nombre de prédicteurs à prendre en compte dans le modèle . La sélection de variables est un point important en pratique pour rendre son modèle plus parcimonieux : plus facile à interpréter et plus stable. En pratique, en chimiométrie, on a plus de variables explicatives que d’échantillons : il faut alors utiliser une autre méthode, une régression PLS par exemple . Dans tous les cas , il faut veiller à valider son modèle à l’aide de jeux tests et/ou d’une méthode de validation croisée afin de pouvoir réutiliser le modèle pour faire de la prédiction.