La Régression Linéaire

18 octobre 2023

Table des matières

1 Introduction

La **Régression Linéaire** est un outil statistique pour établir des prévisions. Son but est d'estimer la relation qu'il existe entre deux ou plusieurs variables : s'il n'y a qu'une variable dite **explicative**, il s'agit de régression linéaire simple. Sinon, on parle de **régression multiple**.

Toutes les régressions ne sont pas linéaires, selon les relations qu'il existe entre les données, on parlera de régressions non linéaires.

Cette méthode doit son nom à Francis Galton (1822 - 1911), mathématicien, statisticien et sociologue, qui fut le premier à l'utiliser, notamment pour prédire la taille d'un enfant à l'âge adulte en fonction de la taille de ses parents.

La Régression Linéaire est aujourd'hui utilisée dans toute sorte de domaines : la prévision commerciale, la finance, le domaine de la santé, dans les études économiques, etc. Il s'agit d'un des algorithmes les plus utilisés en Machine Learning.

2 Notations

2.0.1 Notation simple (ou scalaire)

Pour chaque individu i, la variable expliquée s'écrit comme une fonction linéaire des variables explicatives :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_K x_{i,k} + \varepsilon_i \tag{1}$$

Où :

- y_i est la variable dépendante (la variable que vous essayez de prédire) pour l'individu i. - β_0 est l'intercept, la valeur de y lorsque toutes les variables explicatives sont égales à zéro. - $\beta_1, \beta_2, \ldots, \beta_K$ sont les coefficients de régression, représentant la relation entre la variable dépendante y et les variables explicatives $x_1, x_2, \ldots, x_K$. - $x_{i,1}, x_{i,2}, \ldots, x_{i,K}$ sont les valeurs des variables explicatives correspondantes pour l'individu i. - ε_i est le résidu, l'erreur ou la différence entre la valeur observée y_i et la valeur prédite $\hat{y}_i$ (la valeur obtenue à partir du modèle).

Pour trouver les valeurs des $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ et ε_i, vous utilisez la méthode des moindres carrés ordinaires (MCO) qui vise à minimiser la somme des carrés des résidus. Cette méthode consiste à ajuster le modèle pour qu'il soit aussi proche que possible des données observées.

Les résidus, ε_i, sont les différences entre les valeurs observées de y_i et les valeurs prédites par le modèle $(y_i - \hat{y}_i)$. Ils mesurent l'erreur de prédiction du modèle pour chaque individu de l'échantillon.

2.0.2 Notation matricielle

On a : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$
Avec : i : la i-ème observation de la variable Y, β : les coefficients du modèle, $x_{1,i}$ à $x_{p,i}$ les i-ème observation de la j-ème variable explicative et ε_i l'erreur du modèle.
Cela est équivalent à l'écriture matricielle : $Y = X\beta + \varepsilon$
Soit :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} 1 & x_1 1 & \cdots & x_{1K} \\ 1 & x_2 1 & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Donc :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 1 & \cdots & x_{1K} \\ 1 & x_2 1 & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2.0.3 Explications

Le modèle linéaire est utilisé dans un grand nombre de champs disciplinaires. Il en résulte une grande variété dans la terminologie. Soit le modèle suivant :

$$Y = X\beta + \varepsilon$$

La variable Y est appelée variable expliquée ou variable endogène. Les variables X sont appelées variables explicatives ou variables exogènes. ε est appelé terme d'erreur ou perturbation.

On note généralement $\hat{\beta}$ le vecteur des paramètres estimés. On définit la valeur prédite ou ajustée $\hat{Y} = X\hat{\beta}$ et le résidu comme la différence entre la valeur observée et la valeur prédite :

$$\hat{\varepsilon} = Y - \hat{Y}$$

On définit aussi la somme des carrés des résidus (SCR) comme la somme sur toutes les observations des carrés des résidus :

$$\text{SCR} = \hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

3 Modèle de régression linéaire

3.1 Modèle linéaire simple

Un modèle linéaire simple est un modèle avec **une seule variable explicative**. On a donc deux variables aléatoires, une variable expliquée Y, qui est un scalaire, une variable explicative X, également scalaire. On dispose de n réalisations de ces variables.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2}$$

Dans le modèle linéaire simple, nous pouvons avoir une représentation graphique. L'estimation du modèle linéaire permet de tracer la droite de régression, d'équation $y = \beta_0 + \beta_1 x$. Le paramètre β_0 représente l'ordonnée à l'origine et β_1 le coefficient directeur de la droite.

3.2 Modèle linéaire multiple

Par opposition au modèle de régression linéaire simple, on définit le modèle de régression linéaire multiple comme tout modèle de régression linéaire avec au moins deux variables explicatives.

4 Apllications de la régration Linéaire

4.1 Méthode des moindres carrés ordinaires (MCO)

La méthode des moindres carrés ordinaires (MCO) est l'approche la p lus couramment utilisée pour estimer les coefficients d'un modèle de régression linéaire. Son objectif est de trouver les valeurs des coefficients β qui minimisent la somme des carrés des résidus.

Pour un modèle de régression linéaire simple, la fonction que nous cherchons à minimiser est la somme des carrés des écarts entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$:

$$\mathrm{S} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

Cette équation mesure la "qualité" du modèle en termes de la proximité des prédictions aux données réelles. L'objectif est de trouver les valeurs des coefficients β_0 (intercept) et β_1 (pente) qui minimisent cette somme des carrés.

La solution analytique des MCO pour un modèle de régression linéaire simple est la suivante :

$$\hat{\beta}_1 = \frac{\sum x_i \sum y_i - n \sum x_i y_i}{\left(\sum x_i\right)^2 - n \sum x_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Où $\bar{x}$ est la moyenne empirique des x_i et $\bar{y}$ est la moyenne empirique des y_i.

L'idée fondamentale derrière les MCO est de trouver les coefficients qui minimisent l'erreur de prédiction en ajustant la droite de régression aux données observées. Cette méthode est couramment utilisée pour estimer les coefficients d'un modèle de régression linéaire simple.

Pour un modèle de régression linéaire multiple, la méthode des moindres carrés ordinaires suit le même principe en minimisant la somme des carrés des résidus, mais les calculs sont effectués avec des matrices pour estimer les coefficients β de manière simultanée.

4.2 Exemple de Régression Multiple avec les Moindres Carrés Ordinaires (MCO)

Supposons que nous ayons collecté un ensemble de données comprenant le prix des maisons (Y), la surface (X_1), le nombre de chambres (X_2), et l'emplacement (X_3) pour n maisons. Nous souhaitons ajuster un modèle de régression multiple pour prédire les prix des maisons en fonction de ces variables.

Le modèle de régression multiple est le suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Nous avons les données suivantes :

X_1	X_2	X_3	Y
150	3	1	250
200	4	0	220
120	2	1	150
180	3	1	200
210	4	0	230
140	2	0	160
190	3	1	210
160	3	0	190
130	2	1	170
220	4	1	240

Nous voulons estimer les coefficients $\beta_0, \beta_1, \beta_2$, et β_3 en utilisant la méthode des moindres carrés ordinaires (MCO). Voici comment les estimer :

Les équations de MCO pour les coefficients sont les suivantes :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$
$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$
$$\hat{\beta}_3 = \frac{\sum_{i=1}^{n}(X_{3i} - \bar{X}_3)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{3i} - \bar{X}_3)^2}$$

Où $\bar{Y}$ est la moyenne des prix des maisons, $\bar{X}_1$ est la moyenne de la surface, $\bar{X}_2$ est la moyenne du nombre de chambres, et $\bar{X}_3$ est la moyenne de l'emplacement.

En appliquant ces formules avec les données de l'exemple, vous pouvez estimer les coefficients $\beta_0, \beta_1, \beta_2$, et β_3 pour votre modèle de régression multiple.

Ce modèle vous permettra de prédire les prix des maisons en fonction de la surface, du nombre de chambres et de l'emplacement.

5 Régression linéaire avec des données qualitatives

La régression linéaire est couramment utilisée pour modéliser des relations entre une variable dépendante continue et des variables explicatives continues ou catégorielles. Lorsque vous avez des données qualitatives (catégorielles), vous devez les encoder en variables indicatrices (variables binaires) pour les inclure dans le modèle.

Supposons que vous ayez une variable catégorielle X avec plusieurs niveaux (par exemple, catégories A, B et C). Vous pouvez créer des variables indicatrices pour chacun de ces niveaux, de la manière suivante :

- X_A : Une variable indicatrice pour la catégorie A, prenant la valeur 1 si l'observation appartient à la catégorie A, sinon 0. - X_B : Une variable indicatrice pour la catégorie B, prenant la valeur 1 si l'observation appartient à la catégorie B, sinon 0. - X_C : Une variable indicatrice pour la catégorie C, prenant la valeur 1 si l'observation appartient à la catégorie C, sinon 0.

Ensuite, vous pouvez inclure ces variables indicatrices dans le modèle de régression linéaire comme des variables explicatives. Le modèle serait de la forme :

$$y_i = \beta_0 + \beta_1 X_A + \beta_2 X_B + \beta_3 X_C + \varepsilon_i$$

Chaque coefficient $\beta_1, \beta_2, \beta_3, \ldots$ représente l'effet de chaque catégorie par rapport à une catégorie de référence (généralement laissée de côté). Par exemple, si X_A est 1 et X_B et X_C sont 0, l'effet de la catégorie A est capturé par β_1, et ainsi de suite.

Cela permet de modéliser comment les catégories catégorielles influencent la variable dépendante tout en maintenant un modèle linéaire.