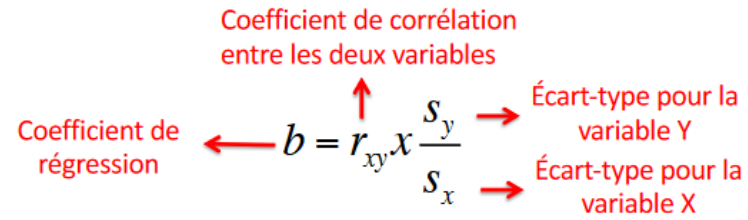
Introduction

La régression linéaire un une formule mathématique qui se sert de la corrélation entre deux variables pour **prédire** la position de la variable y à partir de la position de la variable x. Par exemple, s'il fait 35° Celsius, combien de crème glacée madame D. Queen vendra-t-elle ? Dans une régression linéaire, la variable indépendante devient la **variable prédictrice** et la variable dépendante devient la **variable prédite**. Elles représentent encore la même chose, c'est juste la terminologie qui change.

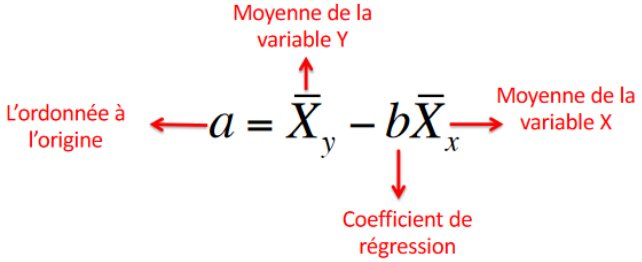
Premièrement, il faut calculer le coefficient de régression. Deuxièmement, il faut calculer l'ordonnée à l'origine (a). Il s'agit de la valeur de y quand x est égale à 0. Troisièmement, il faut calculer la valeur prédite, c'est-à-dire la valeur de y avec notre x donné. Quatrièmement, il faut calculer l'erreur-type d'estimation. Il s'agit de la marge d'erreur de notre modèle. Finalement, on peut calculer une fourchette de valeur prédite, c'est-à-dire une fourchette dans laquelle notre valeur prédite se trouvera. Cette fourchette peut être plus ou moins grande selon la force de notre corrélation. Une corrélation plus forte donnera une fourchette plus petite car on pourra prédire plus précisément où se situera la valeur que l'on veut prédire. Inversement, une corrélation plus faible nous donnera une fourchette plus grande car notre modèle ne nous permettra de prédire de façon plus précise.

Formules

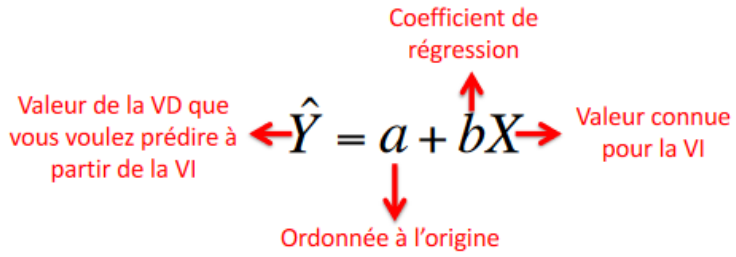
Coefficient de régression



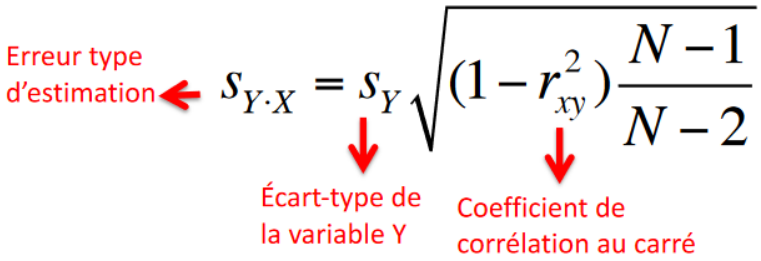
Ordonnée à l'origine



Valeur da la la VD que vous voulez prédire à partir de la VI



Erreur type d'estimation



Mise en situation

Alanis est chanteuse et sort son nouvel album. Alanis est également statisticienne dans son temps libre et elle a établi une corrélation positive entre le nombre d'ironies présentes dans les paroles des chansons de ses albums et la note donnée par les journalistes du magazine Sliding Stone.

La corrélation entre les deux variables est de .80 ; la moyenne de la variable X est 33.43 et son écart-type est de 14.27 ; la moyenne de la variable Y est de 82.14 et son écart-type est de 10.76. Son prochain album contient 42 ironies dans les paroles de ses chansons. Quel note son album risque-t-il d'avoir ?

Albums	Ironies (x)	Note (y)
1) Gentle Little Pill	48	94
2) Confirmed Former Infatuation Junkie	23	65
3) Over Rug Swept	39	93
4) So-Called Order	51	86
5) Odors of Entanglement	12	72
6) Havoc and Bright Nights	24	79
7) Such Pretty Spoons in the Road	37	86

Calculs

Étape 1 : Calculer le coefficient de régression

$$b = r_{xy} * \frac{s_y}{s_x}$$

$$b = 0.8 * \frac{10.76}{14.27} = 0.80 * 0.95 = 0.76$$

Étape 2 : Calcul de l'ordonnée à l'origine

$$a = \underline{X}_y - b\underline{X}_Y$$

$$a = 82.14 - 0.76 * 33.43 = 82.14 - 25.41 = 56.73$$

Étape 3 : Valeur de la VD que vous voulez prédire à partir de la VI

$$\hat{Y} = a + bX$$

$$\hat{Y} = 56.73 + 0.76 * 42 = 56.73 + 31.92 = 88.65$$

Étape 4 : Erreur type d'estimation

$$S_{YX} = S_Y \sqrt{(1 - r_{xy}^2)\frac{N-1}{N-2}}$$

$$S_{YX} = 10.76 \sqrt{(1 - 0.8)\frac{7-1}{7-2}}$$

$$S_{YX} = 10.76 \sqrt{(0.2)\frac{7-1}{7-2}}$$

$$S_{YX} = 10.76 \sqrt{(0.2)\frac{6}{5}}$$

$$S_{YX} = 10.76\sqrt{(0.2)1.2}$$

$$S_{YX} = 10.76\sqrt{0.24}$$

$$S_{YX} = 10.76 * 0.49$$

$$S_{YX} = 5.27$$

Étape 5 : Fourchette de valeur prédite

$$\hat{Y} = \pm S_{YX}$$

$$88.65 - 5.27 = 83.38$$

$$88.65 + 5.27 = 93.92$$

Selon notre modèle, si le prochain album contient 42 ironies, il devrait être noté entre 83.38 et 93.92 (ou 88.65±5.27)

Signification des symbole

X : Valeur

N : Nombre d'observations total

n : Nombre d'observations d'un groupe
- n_1 : nombre d'observations du groupe 1 ;
- n_2 : nombre d'observations du groupe 2 ;
- etc.

K : Nombre de groupes

Σ : Somme

$\bar{X}$: Moyenne d'un échantillon

μ : Moyenne d'une population

s : Écart-type d'un échantillon

σ : Écart-type d'une population

s^2 : Variance d'un échantillon

σ^2 : Variance d'une population

χ^2 : Khi-carré

A: Fréquence attendue

O: Fréquence observé

L: Ligne

C: Colonne

t : Statistique t (ou score t dans le cas d'une corrélation/régression linéaire)

r : Coefficient de Pearson

Z : Score Z

b : Coefficient de régression

a : Ordonnée à l'origine

$\hat{Y}$: Valeur de la VD qu'on veut prédire l'aide la VI

F : Statistique F

Ouvrages de référence

- Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics: North American Edition* (5th ed.). Sage Edge
- Howell. D.C. (2008). *Méthodes statistiques en sciences humaines* (M. Rogier, V. Yzerbyt, & Y. Bestgen, Trans.). (6th ed.). De boeck. (Original work published 2008)
- Tabachnick, B. G., & Fidell, L. S. (2021). *Using Multivariate Statistics* (7th ed.). Pearson