

Série 1

Régression linéaire

Exercice 1 : Estimation des coefficients de régression

Un chercheur observe la relation entre le nombre d'heures d'étude (x) et la note obtenue (y) par des étudiants. Les données suivantes sont disponibles :

x	y
1	2
2	3
3	5
4	4
5	6

- 1) Proposez une relation linéaire entre x et y sous la forme $y = \beta_0 + \beta_1 x$
- 2) Utilisez la méthode des moindres carrés pour calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
- 3) Écrivez l'équation de la droite de régression.
- 4) Prédisez la note pour un étudiant qui a étudié 6 heures.
- 5) Calculez l'erreur quadratique moyenne (MSE) pour ce modèle.

Exercice 2 : Implémentation Python

Importez une bibliothèque de machine learning (par exemple, `scikit-learn`) et chargez les données suivantes :

$$X = \{1, 2, 3, 4, 5, 6\}, \quad y = \{2, 4, 5, 4, 5, 7\}.$$

1. Implémentez une régression linéaire pour ajuster les données.
2. Affichez les coefficients β_0 (intercept) et β_1 (slope).
3. Tracez les points de données (x, y) ainsi que la droite de régression sur un graphique.
4. Utilisez le modèle pour prédire la valeur de y lorsque $x = 7$.

Exercice 3 : Évaluation du modèle

Les données suivantes représentent la taille (x) et le poids (y) de différentes personnes :

x (taille, en cm)	150	160	170	180	190
y (poids, en kg)	50	60	65	70	80

1. Ajustez une droite de régression linéaire aux données.
2. Calculez le coefficient de détermination R^2 .
3. Expliquez si le modèle est un bon ajustement pour ces données.
4. Ajoutez un nouveau point ($x = 200, y = 90$) aux données. Réajustez le modèle et comparez les nouveaux paramètres et le R^2 .

Exercice 4 : Données bruitées

1. Générez des données artificielles avec du bruit en suivant la relation :

$$y = 2x + 1 + \varepsilon,$$

où ε est une erreur générée aléatoirement selon une loi normale $\mathcal{N}(0, 1)$.

2. Générez 50 points x aléatoires dans l'intervalle $[0, 10]$.
3. Ajustez une régression linéaire à ces données et tracez le graphique.
4. Analysez les résidus (différence entre les y observés et les y prévus).
5. Expliquez l'impact du bruit sur la précision de la régression.

Exercice 5 : Régression non linéaire

Vous disposez des données suivantes :

x	1	2	3	4	5	6
y	1	4	9	16	25	36

1. Montrez que les données ne peuvent pas être bien ajustées par un modèle linéaire.
2. Proposez une transformation linéarisante, par exemple, $y' = \sqrt{y}$, et réessayez une régression linéaire.
3. Comparez les modèles avant et après transformation.

Exercice 6 : Régression multiple (Avancé)

Vous disposez des données suivantes pour modéliser le prix d'une maison (y) en fonction de sa surface (x_1) et du nombre de chambres (x_2) :

Surface (x_1 , en m ²)	50	60	70	80	90
Chambres (x_2)	2	2	3	3	4
Prix (y , en k€)	200	220	250	280	310

1. Formulez le modèle linéaire sous la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

2. Utilisez la méthode des moindres carrés pour estimer les coefficients β_0 , β_1 , et β_2 .
3. Prédisez le prix d'une maison de 85 m² avec 3 chambres.
4. Comparez vos résultats à ceux obtenus en utilisant une bibliothèque Python.

Exercice 7 : Régression et biais/variance

1. Générez 100 points selon la relation $y = 3x + 2$ avec différents niveaux de bruit (par exemple, bruit faible, bruit moyen, bruit élevé).
2. Ajustez un modèle linéaire pour chaque jeu de données.
3. Tracez les données, les modèles ajustés, et discutez de l'impact du bruit en termes de biais et variance.
4. Expliquez comment la taille des données affecte la performance de la régression.

Exercice 8 : Interprétation des résidus

On considère un jeu de données simple pour effectuer une régression linéaire et analyser les résidus.

1. Ajuster un modèle de régression linéaire sur les données suivantes :

$$x = [1, 2, 3, 4, 5], \quad y = [1.5, 2.5, 3.5, 4.5, 5.5].$$

2. Calculer les résidus, définis comme la différence entre les valeurs observées y et les valeurs prédites \hat{y} , après l'entraînement du modèle.
3. Tracer un graphique des résidus en fonction des valeurs de x et vérifiez si les résidus sont distribués aléatoirement autour de 0..
4. Analyser le graphique des résidus pour évaluer la qualité de l'ajustement du modèle.