Régression linéaire

Deux méthodes pour trouver la même droite : calcul exact ou descente de gradient

À savoir avant de commencer

Ce document réutilise la descente de gradient (document 0) et la compare à une méthode de calcul exact, la méthode des moindres carrés ordinaires (MCO). C'est l'occasion de comprendre pourquoi l'IA moderne utilise presque toujours la descente de gradient, même quand une formule exacte existe.

Niveau : Progressif : Seconde → Terminale (spécialité et option mathématiques complémentaires, précisé à chaque partie)

Introduction : prédire un nombre, pas une catégorie

Contrairement au document 3 (classifier des fruits en deux catégories), on s'intéresse ici à un autre type de tâche très fréquent en intelligence artificielle : prédire une valeur numérique continue à partir d'une autre. On parle de régression.

Exemple : peut-on prédire le score qu'obtiendra un élève à un test, à partir du nombre d'heures qu'il a passées à réviser ?

On cherche une droite y = ax + b qui s'ajuste au mieux à un nuage de points.

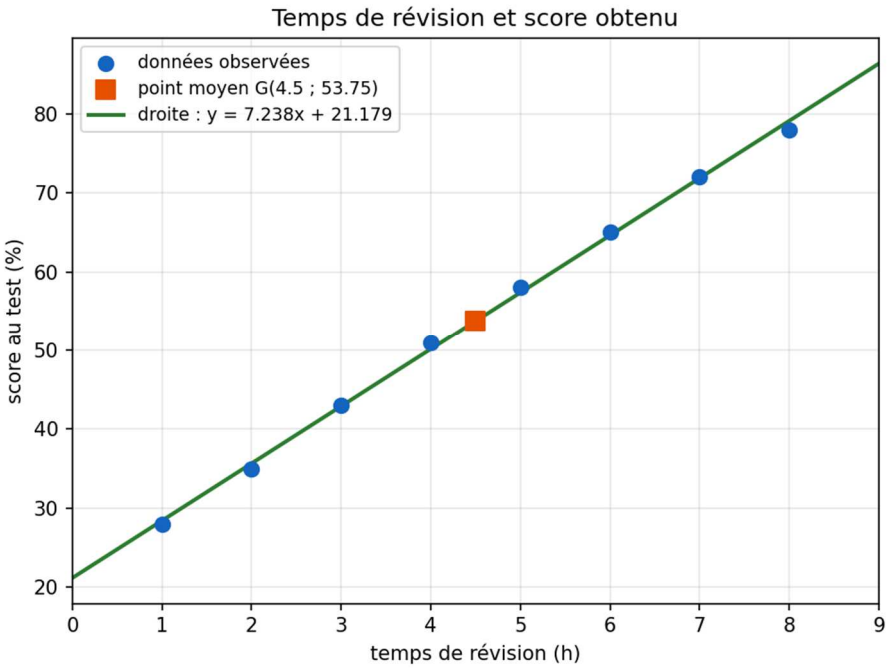
La question que ce document explore est double : comment trouver cette droite (deux méthodes possibles), et laquelle de ces deux méthodes une IA utilise-t-elle en pratique ?

Partie A — Un nuage de points bien aligné

Niveau : Seconde

On a demandé à 8 élèves combien d'heures ils ont révisé avant un test, et quel score ils ont obtenu :

Temps de révision (h)	1	2	3	4	5	6	7	8
Score obtenu (%)	28	35	43	51	58	65	72	78



1. D'après le nuage de points, un ajustement affine (une droite) te semble-t-il pertinent pour décrire la relation entre les deux grandeurs ?
2. Calcule les coordonnées du point moyen G($\bar{x}$; $\bar{y}$) de ce nuage de points.
3. Plusieurs droites pourraient passer visuellement « au mieux » parmi les points. Comment choisir, parmi toutes ces droites, la meilleure possible ? (Tu peux relire le document 0 pour t'en inspirer.)

Partie B — La méthode des moindres carrés (calcul exact)

Niveau : Terminale — option mathématiques complémentaires

Point de vigilance

La formule explicite de la droite de régression (ci-dessous) appartient au programme de l'option mathématiques complémentaires de terminale, et non à celui de la spécialité mathématiques (où l'ajustement affine se fait « à l'œil » ou à la calculatrice, sans redémontrer la formule). Les élèves de spécialité peuvent tout à fait suivre cette partie : elle ne fait qu'appliquer des outils déjà connus (moyenne, sommes).

Il existe une formule directe, qui donne exactement les coefficients a et b de la droite qui minimise la somme des carrés des écarts entre les points et la droite (d'où le nom « moindres carrés ») :

Formules de la droite de régression (MCO)

$$a = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \,/\, \Sigma(x_i - \bar{x})^2 \qquad b = \bar{y} - a \times \bar{x}$$

Complète le tableau suivant pour préparer le calcul ($\bar{x} = 4{,}5$; $\bar{y} = 53{,}75$) :

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	28	?	?	?	?
2	35	?	?	?	?
3	43	?	?	?	?
4	51	?	?	?	?
5	58	?	?	?	?
6	65	?	?	?	?
7	72	?	?	?	?
8	78	?	?	?	?

4. En sommant les deux dernières colonnes, calcule $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ et $\Sigma(x_i - \bar{x})^2$.
5. En déduire les valeurs exactes de a puis de b (arrondis à 10^{-3} près).
6. Écris l'équation de la droite de régression obtenue.
7. Utilise cette équation pour prédire le score d'un élève ayant révisé 5h30, puis celui d'un élève ayant révisé 12h. Ce deuxième résultat te semble-t-il réaliste ? Pourquoi ?
8. Interprète, dans le contexte de l'énoncé, le coefficient directeur a de la droite.

Partie C — Retrouver la même droite par descente de gradient

Niveau : Terminale

On peut aussi retrouver cette droite sans formule exacte, en réutilisant la descente de gradient du document 0. On définit la fonction de coût (erreur quadratique moyenne) :

Fonction de coût

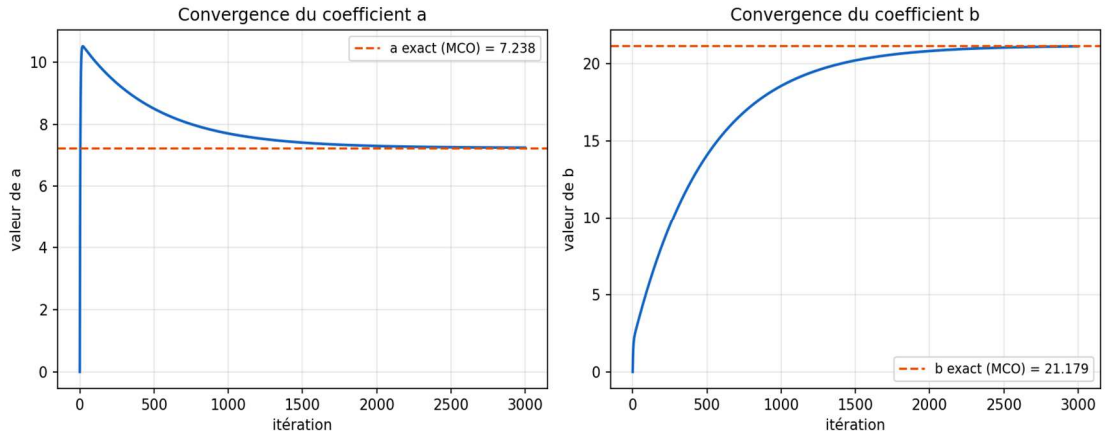
$$J(a, b) = (1/2n) \times \Sigma_i \, (a\, x_i + b - y_i)^2$$

En dérivant (comme en document 0, mais maintenant à deux paramètres) et en moyennant sur les n exemples, on obtient : $\partial J/\partial a = (1/n)\Sigma_i(\hat{y}_i - y_i)x_i$ et $\partial J/\partial b = (1/n)\Sigma_i(\hat{y}_i - y_i)$, avec $\hat{y}_i = a\, x_i + b$.

On part de $a = 0$ et $b = 0$, avec un pas d'apprentissage $\eta = 0{,}01$ (volontairement petit, on verra pourquoi à la fin de cette partie).

9. Avec $a = 0$ et $b = 0$, on a $\hat{y}_i = 0$ pour chaque élève, donc $\hat{y}_i - y_i = -y_i$. Calcule $\partial J/\partial a = -(1/8)\Sigma x_i y_i$ et $\partial J/\partial b = -(1/8)\Sigma$□ (tu peux réutiliser les données de la partie B ; $\Sigma x_i y_i = 2239$).
10. Effectue la mise à jour de a et b avec $\eta = 0{,}01$. Compare ces valeurs à celles obtenues à la partie B : sont-elles déjà proches ?

Le graphique ci-dessous montre, calculée par ordinateur, l'évolution de a et de b au fil de 3 000 itérations de descente de gradient, comparée à la valeur exacte trouvée par la méthode des moindres carrés en partie B :



11. Que remarques-tu sur les valeurs de a et de b après plusieurs milliers d'itérations, comparées à celles de la partie B ?
12. Sur le graphique de a, on observe que la valeur « dépasse » légèrement la cible avant de redescendre. Ce comportement te rappelle-t-il quelque chose vu dans le document 0 ?
13. Pourquoi a-t-on choisi ici un pas d'apprentissage η aussi petit (0,01) ? Que se passerait-il, à ton avis, avec un pas beaucoup plus grand, sachant que les données (temps, score) prennent des valeurs assez élevées (jusqu'à x=8 et y=78) ?

Partie D — Comparer les deux méthodes

Niveau : Terminale

	Moindres carrés (MCO)	Descente de gradient
Principe	Formule directe, calcul exact en une seule étape	Méthode itérative, on s'approche progressivement de la solution
Résultat	Exact (aux arrondis près)	Approché, d'autant plus précis qu'on itère longtemps
Paramètres à choisir	Aucun	Pas d'apprentissage η, nombre d'itérations
Avec beaucoup de données (millions de points)	Coûteux, voire impossible en pratique	Reste utilisable
Avec une fonction de coût plus complexe (non quadratique)	Formule exacte généralement introuvable	Fonctionne quand même

14. D'après ce tableau, dans quel cas la méthode des moindres carrés est-elle clairement préférable à la descente de gradient ?
15. Un ingénieur doit entraîner un réseau de neurones avec 50 millions de paramètres, sur un cœur de métier où la fonction de coût n'est pas une simple parabole. Quelle méthode va-t-il utiliser ? Justifie.

Partie E — Pourquoi l'IA moderne utilise (presque) toujours la descente de gradient

Niveau : Tous niveaux

Pour une régression linéaire à une seule variable comme celle de ce document, la méthode des moindres carrés est en réalité la meilleure option : elle est exacte, rapide, et ne nécessite aucun réglage. La descente de gradient n'a ici qu'un intérêt pédagogique — montrer qu'elle retrouve bien le même résultat.

Mais dès que le modèle se complexifie (des milliers de variables, des fonctions de coût non quadratiques, des réseaux de neurones à plusieurs couches), aucune formule exacte n'existe en général.

La descente de gradient devient alors la seule option praticable : c'est pour cette raison qu'elle est la méthode d'apprentissage centrale de la quasi-totalité des IA actuelles, y compris les modèles de langage comme Claude — même si, pour un problème aussi simple qu'une droite de régression, une formule exacte suffit largement.

Ce que ce modèle simplifie, et ce que fait réellement une IA comme Claude

Ce que ce modèle simplifie	Ce que fait réellement un modèle comme Claude
Une seule variable d'entrée (le temps de révision) et une relation linéaire simple. Une fonction de coût quadratique, pour laquelle une formule exacte (MCO) existe. Une descente de gradient présentée ici comme une alternative parmi d'autres.	Des milliers à des milliards de variables d'entrée (les coordonnées des vecteurs embeddings — document 2), combinées de façon fortement non linéaire. Une fonction de coût si complexe (des milliards de paramètres, de nombreuses couches non linéaires) qu'aucune formule exacte n'existe : on ne peut même pas être certain de trouver LE minimum global, seulement un bon minimum local. La descente de gradient (et ses variantes plus sophistiquées, comme Adam, qui adaptent automatiquement le pas d'apprentissage) comme unique méthode praticable à cette échelle.

L'idée centrale — minimiser une fonction de coût qui mesure l'écart entre prédictions et réalité — reste, elle, exactement ce qui se passe à l'intérieur de tout modèle d'IA, qu'une solution exacte existe ou non.

Exercices d'entraînement

On a mesuré, chez 5 personnes, leur nombre d'heures de sommeil la nuit précédente x, et un score de concentration y obtenu à un test le lendemain matin :

Sommeil (h)	4	5	6	7	8
Score de concentration	40	55	65	72	80

Exercice 1 — Point moyen et ajustement

Niveau : Seconde

16. Calcule les coordonnées du point moyen G($\bar{x}$; $\bar{y}$) de ce nuage de points.
17. Un ajustement affine te semble-t-il pertinent ici ? Justifie à partir de la forme du nuage de points (tu peux le représenter rapidement).

Exercice 2 — Calcul de la droite de régression (MCO)

Niveau : Terminale — option mathématiques complémentaires

On donne $\bar{x} = 6$ et $\bar{y} = 62{,}4$.

18. Complète un tableau $(x_i - \bar{x})$, $(y_i - \bar{y})$, produits et carrés, comme en partie B, pour calculer $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ et $\Sigma(x_i - \bar{x})^2$.
19. En déduire les coefficients a et b de la droite de régression (arrondis à 10^{-1} près).
20. Utilise cette droite pour prédire le score de concentration d'une personne ayant dormi 10 heures. Ce résultat est-il réaliste ?

Exercice 3 — Une itération de descente de gradient

Niveau : Terminale

On applique la descente de gradient à ce même jeu de données, en partant de $a = 0$ et $b = 0$, avec $\eta = 0{,}01$. On rappelle que $\Sigma x_i y_i = 40\times4+55\times5+65\times6+72\times7+80\times8$ (à calculer), et $\Sigma y_i = 40+55+65+72+80 = 312$.

21. Calcule $\partial J/\partial a$ et $\partial J/\partial b$ avec a=0, b=0 (comme en partie C.1).
22. Effectue une itération de mise à jour de a et b.
23. Compare ce résultat à la valeur exacte de a et b trouvée à l'exercice 2. Une seule itération suffit-elle à s'en approcher ?

Corrigé

Partie A

A.1. Oui : les points semblent globalement alignés selon une tendance croissante, un ajustement affine est pertinent.

A.2. $\bar{x} = (1+2+3+4+5+6+7+8)/8 = 36/8 = 4{,}5$. $\bar{y} = (28+35+43+51+58+65+72+78)/8 = 430/8 = 53{,}75$. G(4,5 ; 53,75).

A.3. On choisit la droite qui minimise une fonction de coût mesurant l'écart entre les points et la droite (par exemple la somme des carrés des écarts) — exactement le même principe que la descente de gradient du document 0, mais ici appliqué à deux paramètres a et b au lieu d'un seul x.

Partie B

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	produit	carré
1	28	−3,5	−25,75	90,125	12,25
2	35	−2,5	−18,75	46,875	6,25
3	43	−1,5	−10,75	16,125	2,25
4	51	−0,5	−2,75	1,375	0,25
5	58	0,5	4,25	2,125	0,25
6	65	1,5	11,25	16,875	2,25
7	72	2,5	18,25	45,625	6,25
8	78	3,5	24,25	84,875	12,25

B.1. $\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 90{,}125+46{,}875+16{,}125+1{,}375+2{,}125+16{,}875+45{,}625+84{,}875 = 304$. $\Sigma(x_i-\bar{x})^2 = 12{,}25+6{,}25+2{,}25+0{,}25+0{,}25+2{,}25+6{,}25+12{,}25 = 42$.

B.2. $a = 304/42 \approx 7{,}238$. $b = 53{,}75 - 7{,}238\times4{,}5 \approx 53{,}75 - 32{,}571 \approx 21{,}179$.

B.3. $y \approx 7{,}238x + 21{,}179$.

B.4. Pour x=5,5 : $y \approx 7{,}238\times5{,}5+21{,}179 \approx 39{,}81+21{,}18 \approx 60{,}99$, soit un score prévu d'environ 61 %. Pour x=12 : $y \approx 7{,}238\times12+21{,}179 \approx 86{,}86+21{,}18 \approx 108{,}03$, soit plus de 100 % : c'est irréaliste (un score ne peut pas dépasser 100 %). Cela montre le danger d'extrapoler un modèle linéaire en dehors de la plage de données observées (ici, 1 à 8 heures).

B.5. Le coefficient $a \approx 7{,}238$ signifie qu'en moyenne, chaque heure supplémentaire de révision est associée à une augmentation d'environ 7,24 points de score.

Partie C

C.1. $\partial J/\partial a = -(1/8)\times2239 = -279{,}875$. $\partial J/\partial b = -(1/8)\times430 = -53{,}75$.

C.2. $a_1 = 0 - 0{,}01\times(-279{,}875) = 2{,}799$. $b_1 = 0 - 0{,}01\times(-53{,}75) = 0{,}538$. Ces valeurs sont encore loin de $a\approx7{,}238$ et $b\approx21{,}179$: une seule itération ne suffit pas.

C.3. Après plusieurs milliers d'itérations, les valeurs de a et de b obtenues par descente de gradient deviennent quasiment identiques à celles trouvées exactement par la méthode des moindres carrés en partie B (aux petites imprécisions numériques près).

C.4. Oui : ce comportement de « dépassement puis correction » rappelle exactement ce qui a été observé dans le document 0 (partie D) avec un pas d'apprentissage trop grand par rapport à la pente locale : la suite peut temporairement s'éloigner avant de se stabiliser.

C.5. Un pas d'apprentissage aussi petit est nécessaire car les données ne sont pas normalisées : x et y prennent des valeurs assez grandes (jusqu'à 8 et 78), ce qui rend les gradients eux-mêmes assez grands (ici environ −280 et −54). Avec un pas plus grand, les mises à jour deviendraient énormes dès la première itération, ce qui ferait diverger la suite (a et b partiraient vers l'infini), exactement comme dans l'exercice 2 du document 0 avec un pas trop grand.

Partie D

D.1. La méthode des moindres carrés est clairement préférable quand on dispose d'une formule exacte applicable (régression linéaire simple, quantité raisonnable de données) : elle est plus rapide, plus précise, et ne nécessite aucun réglage de paramètre.

D.2. Il utilisera la descente de gradient (ou une de ses variantes), car aucune formule exacte n'existe pour une fonction de coût aussi complexe avec autant de paramètres : c'est la seule méthode praticable dans ce cas.

Exercice 1

1. $\bar{x} = (4+5+6+7+8)/5 = 30/5 = 6$. $\bar{y} = (40+55+65+72+80)/5 = 312/5 = 62{,}4$. G(6 ; 62,4).

2. Le nuage de points est croissant et globalement bien aligné : un ajustement affine est pertinent.

Exercice 2

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	produit	carré
4	40	−2	−22,4	44,8	4
5	55	−1	−7,4	7,4	1
6	65	0	2,6	0	0
7	72	1	9,6	9,6	1
8	80	2	17,6	35,2	4

1. $\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 44,8+7,4+0+9,6+35,2 = 97$. $\Sigma(x_i-\bar{x})^2 = 4+1+0+1+4 = 10$.

2. $a = 97/10 = 9,7$. $b = 62,4 - 9,7\times6 = 62,4 - 58,2 = 4,2$.

3. Pour $x=10$: $y = 9,7\times10+4,2 = 101$, soit plus de 100 % : irréaliste (extrapolation en dehors de la plage observée de 4 à 8 heures).

Exercice 3

$\Sigma x_i y_i = 40\times4+55\times5+65\times6+72\times7+80\times8 = 160+275+390+504+640 = 1969$.

1. $\partial J/\partial a = -(1/5)\times1969 = -393,8$. $\partial J/\partial b = -(1/5)\times312 = -62,4$.

2. $a_1 = 0 - 0,01\times(-393,8) = 3,938$. $b_1 = 0 - 0,01\times(-62,4) = 0,624$.

3. Ces valeurs ($a\approx3,938$; $b\approx0,624$) sont encore loin des valeurs exactes trouvées à l'exercice 2 ($a=9,7$; $b=4,2$) : une seule itération ne suffit pas, il en faudrait plusieurs milliers pour converger, comme observé en partie C.

Mathématiques du programme mobilisées dans ce document

Seconde

Nuages de points, point moyen, ajustement affine (« à l'œil » ou à la calculatrice).

Moyenne d'une série statistique.

Terminale — option mathématiques complémentaires

Droite de régression par la méthode des moindres carrés (formule explicite des coefficients a et b) : ce contenu n'appartient pas à la spécialité mathématiques (voir partie B).

Interprétation d'un coefficient directeur et d'une ordonnée à l'origine dans un contexte concret.

Terminale (spécialité mathématiques)

Fonction de coût quadratique à deux variables, dérivées partielles (lien avec le document 0).

Descente de gradient appliquée à deux paramètres simultanément.

Analyse critique d'un modèle (danger de l'extrapolation en dehors du domaine observé).