

# INTRODUCTION AUX OUTILS DE L'INTELLIGENCE ARTIFICIELLE

PARTIE 1 : INTRODUCTION À L'INTELLIGENCE ARTIFICIELLE

PARTIE 2 : RÉGRESSION LINÉAIRE SIMPLE

V. Monbet

<sup>1</sup>Université de Rennes/UFR Mathématiques

# Outline

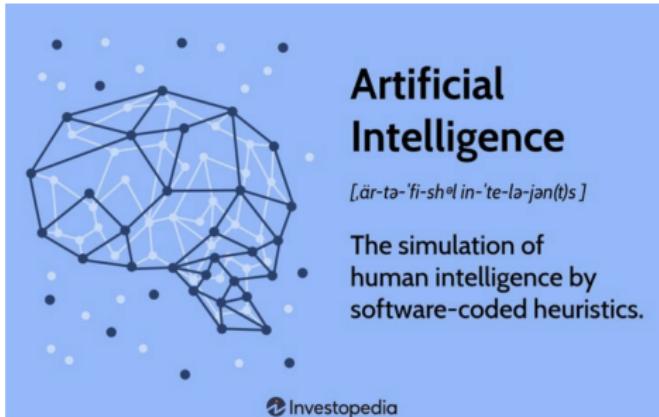
Quelques mots sur l'intelligence artificielle

Régression linéaire univariée



# WHAT IS A.I.?

Lorsque la plupart des gens entendent le terme "intelligence artificielle", la première chose qui leur vient à l'esprit, ce sont les robots. C'est parce que les films et les romans à gros budget racontent l'histoire de machines à l'apparence humaine qui font des ravages sur Terre. Mais rien n'est plus faux.



## Artificial Intelligence

[är-tə-'fi-shəl in-'te-lə-jən(t)s]

The simulation of  
human intelligence by  
software-coded heuristics.

 Investopedia

L'intelligence artificielle repose sur le principe selon lequel l'intelligence humaine peut être définie de manière à ce qu'une machine puisse facilement l'imiter et exécuter des tâches, des plus simples aux plus complexes.

# Les machines peuvent-elles être aussi créatives que les humains ?



Mario Klingemann's "Love Wound" (2021), created using his AI algorithm [Botto](#)

OpenAI, la société à l'origine de ChatGPT, répond par l'affirmative avec la sortie récente de GPT-4, son système le plus avancé à ce jour.

Outre le contenu écrit, les outils basés sur l'IA démontrent leurs prouesses dans d'autres domaines créatifs tels que la musique, les arts visuels, la photographie et la vidéographie.

Jukebox, le cousin moins connu de ChatGPT, génère de la musique (sous forme d'audio brut) lorsqu'on lui fournit des données telles que l'artiste et le genre. De son côté, l'artiste allemand Mario Klingemann est à la tête du mouvement artistique de l'IAs.

## Autres exemples

- ▶ **Assistants virtuels intelligents**

Les assistants virtuels intelligents tels que Siri, Alexa et Google Assistant font désormais partie intégrante de notre vie quotidienne. Ces outils alimentés par des applications d'IA nous permettent d'effectuer facilement des tâches en gardant les mains libres.

- ▶ **Diagnostic précis et plans de traitement efficaces**

L'IA a un impact considérable sur le secteur de la santé en améliorant la précision des diagnostics et en optimisant les plans de traitement. Grâce à leur capacité à analyser de grandes quantités de données médicales, les algorithmes d'IA peuvent fournir aux médecins des suggestions plus précises et plus exactes pour les soins aux patients.

## Autres exemples

### ► **Contrôle qualité**

Les algorithmes d'IA sont utilisés pour améliorer le contrôle de la qualité dans l'industrie manufacturière. Les algorithmes sont formés pour identifier des modèles dans les données de qualité et peuvent signaler des problèmes potentiels avant qu'ils ne se produisent. Cela permet d'améliorer la qualité des produits et de réduire le risque de rappels de produits coûteux et d'autres problèmes de qualité.

### ► **Détection de la fraude**

La détection des fraudes est une application cruciale de l'intelligence artificielle dans la finance. Elle peut analyser de grandes quantités de données financières et détecter des cas de fraude potentiels, réduisant ainsi le risque de fraude et permettant aux institutions financières de répondre rapidement aux menaces potentielles.

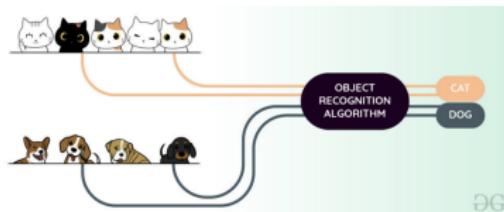
## Comment ça marche ?

En général, les systèmes d'IA fonctionnent en ingérant de grandes quantités de données d'apprentissage étiquetées, en analysant les données à la recherche de corrélations et de modèles, et en utilisant ces modèles pour faire des prédictions sur les états futurs.

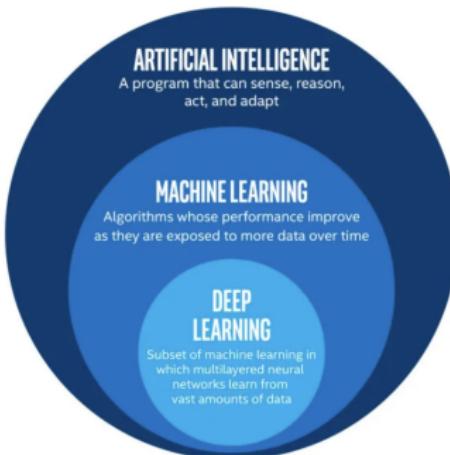
C'est ainsi qu'un chatbot nourri d'exemples de texte peut apprendre à générer des échanges réalistes avec des personnes,



ou qu'un outil de reconnaissance d'images peut apprendre à identifier et à décrire des objets dans des images en examinant des millions d'exemples.



## Parmi les outils clé de l'IA

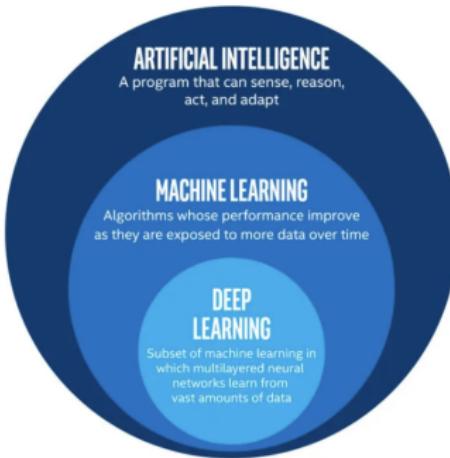


Source [medium.com](#)

### ► Machine Learning

L'apprentissage automatique est un sous domaine d'étude de l'intelligence artificielle qui s'intéresse au développement et à l'étude d'algorithmes statistiques capables d'apprendre à partir de données et de généraliser à des données inédites.

## Parmi les outils clé de l'IA



Source [medium.com](https://medium.com)

### ► Deep Learning

Le Deep Learning est une catégorie de machine learning qui exploite des **réseaux de neurones** artificiels, inspirés directement du fonctionnement des neurones humains, pour assimiler de nouvelles connaissances. Ces réseaux sont constitués de plusieurs couches de neurones artificiels connectés entre eux. Plus le nombre de couches est élevé, plus le réseau est profond et puissant.

## Les mathématiques des réseaux de neurones

Dans la suite de ce cours, nous allons nous intéresser principalement aux réseaux de neurones.

Les outils mathématiques des réseaux de neurones sont

- ▶ Algèbre linéaire : calcul matriciel
- ▶ Optimisation : calcul de dérivées, algorithmes d'optimisation numérique
- ▶ Probabilité et statistique : estimation, prise en compte des incertitudes, validation des outils développés

# Introduction aux réseaux de neurones

## Plan du cours

1. Régression linéaire univariée puis multivariée
2. Optimisation par la méthode de descente du gradient, en dimension un puis en dimension quelconque
3. Réseaux de neurones (profonds) complètement connectés en régression

# Outline

Quelques mots sur l'intelligence artificielle

Régression linéaire univariée

Introduction

Modélisation

Estimation (*Learning*) par la méthode des moindres carrés

# Outline

## Régression linéaire univariée

Introduction

Modélisation

Estimation (*Learning*) par la méthode des moindres carrés

## Introduction

Le but de la régression linéaire simple est d'expliquer et/ou prédire une variable quantitative à partir d'une autre variable (quantitative).

### Exemples

- ▶ Expliquer le prix d'un appartement en fonction de la superficie
- ▶ Prédire la hauteur d'un arbre à partir de sa circonférence
- ▶ Prédire les variations des ventes d'un produit en fonction de l'investissement publicitaire
- ▶ Expliquer ou prédire un niveau de pollution à partir d'une variable météorologique

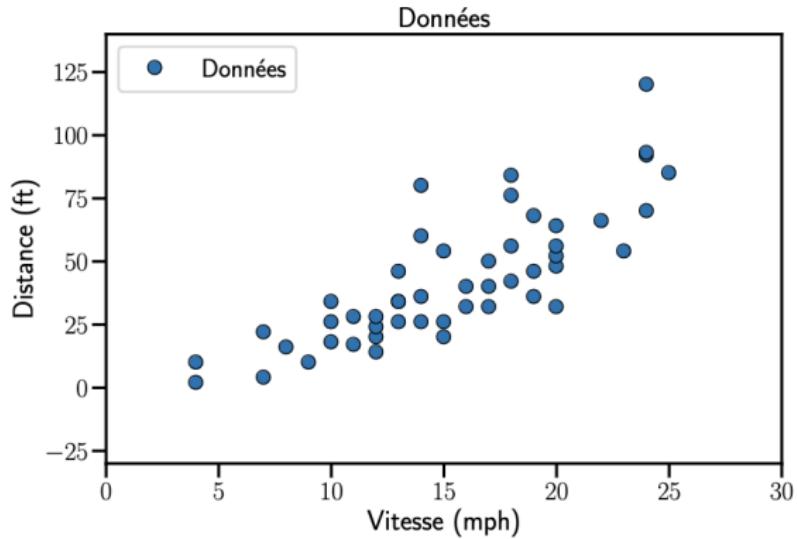
Données : deux variables mesurées / expérience

Étudier :

1. si les variables sont (linéairement) liées
2. la force du lien
3. si la variable d'intérêt peut être prédite en observant uniquement l'autre variable (la variable explicative)

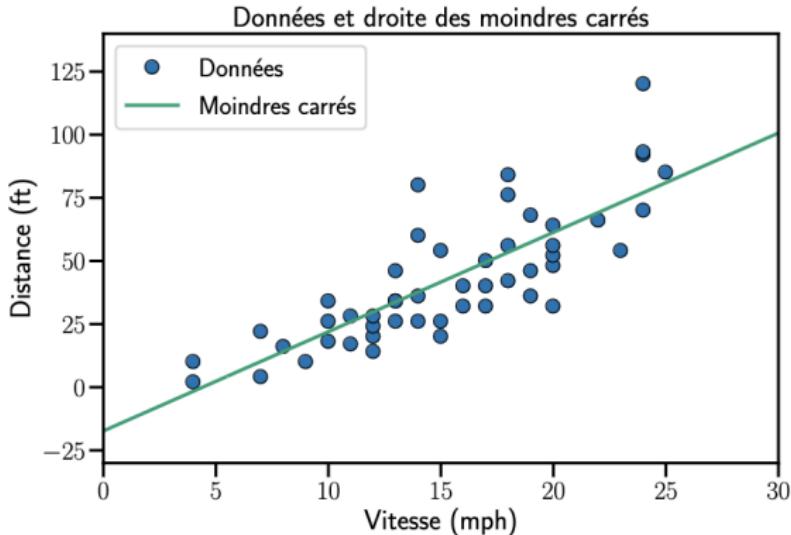
Exemple : vitesse et distance de freinage de voitures

n = 50 mesures



Source : <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html>

Exemple : vitesse et distance de freinage de voitures  
n = 50 mesures



Question : comment "trouver" (on dira **estimer**) la droite de régression ?

# Outline

## Régression linéaire univariée

Introduction

Modélisation

Estimation (*Learning*) par la méthode des moindres carrés

## Modélisation et notations

Observations :  $(x_i, y_i)$  pour  $i = 1, \dots, n$

Hypothèse de modèle linéaire ou **régression linéaire** :

$$y_i \simeq \beta_0^* + \beta_1^* x_i \quad \text{pour tout } i = 1, \dots, n$$

- ▶  $\beta_0^*$  : ordonnée à l'origine (inconnue)
- ▶  $\beta_1^*$  : coefficient directeur (inconnu)

Rem :  $\beta_0^*$  et  $\beta_1^*$  sont les paramètres inconnus qu'on va chercher à estimer ou "apprendre".

### Définitions

- ▶  $y$  est une **variable à expliquer** ou **variable cible**
- ▶  $x$  est une **variable explicative** ou **covariable**

## Interprétation des notations

Exemple : dataset *cars*

- ▶  $n$  = 50, nombre d'observations (ou taille de l'échantillon)
- ▶  $y_i$  : temps de freinage de la voiture  $i$
- ▶  $x_i$  : vitesse de la voiture  $i$
- ▶  $y$  : la variable à prédire est le temps de freinage
- ▶  $x$  : la variable explicative est la vitesse

L'hypothèse de modèle linéaire :

ici cela revient à postuler que le temps de freinage d'une voiture est **proportionnel** à sa vitesse (!)

## Modélisation II

### Modèle probabiliste<sup>1</sup>

$$y_i = \beta_0^* + \beta_1^* x_i + \epsilon_i$$

avec  $\epsilon_i$  une variable aléatoire de moyenne nulle et de variance  $\sigma^2$

### Interprétation

Les  $\epsilon_i = y_i - \beta_0^* - \beta_1^* x_i$  représentent les erreurs entre le modèle théorique et les observations.

L'aspect aléatoire peut provenir du bruit de mesure, du bruit de transmission, de la variabilité d'échantillonage, etc.

---

1. on donne ici un sens au symbole  $\simeq$

# Modélisation III

## Modèle probabiliste

$$y_i = \beta_0^* + \beta_1^* x_i + \epsilon_i$$

### Définitions

On appelle

- ▶  $\beta_0$  l'**ordonnée à l'origine** (intercept ou bias)
- ▶  $\beta_1$  la **pente** (slope ou weight)

**Objectif 1** (*learning*) Estimer  $\beta_0^*$  et  $\beta_1^*$  (inconnus) par des quantités  $\hat{\beta}_0$  et  $\hat{\beta}_1$  dépendant des observations  $(x_i, y_i)$  pour  $i = 1 \dots, n$ .

**Objectif 2** (*prediction*) Prédire pour un nouveau point  $x_{n+1}$  la valeur  $y_{n+1}$  qui n'est pas observée par  $y_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

**Objectif 3** (*validation*)

# Outline

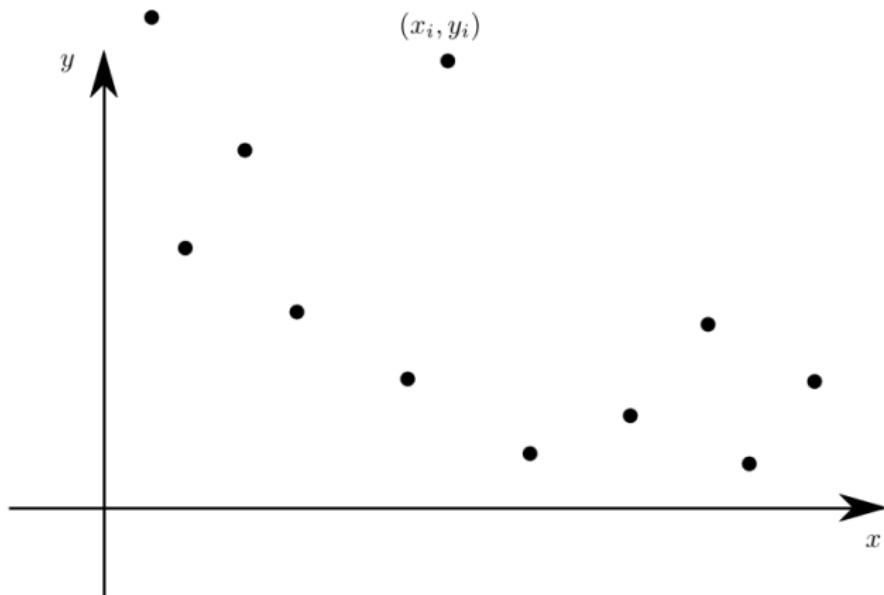
## Régression linéaire univariée

Introduction

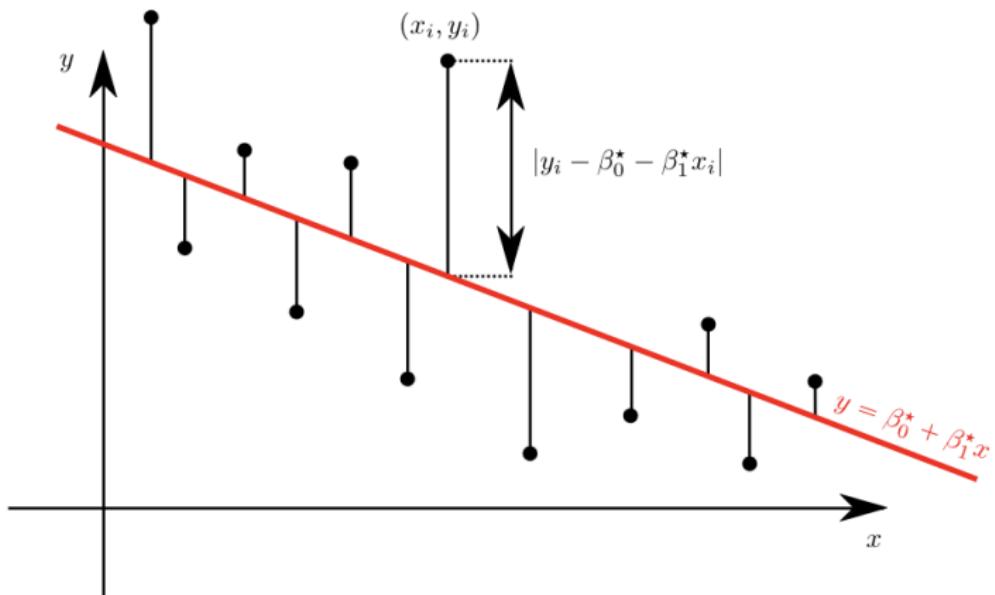
Modélisation

Estimation (*Learning*) par la méthode des moindres carrés

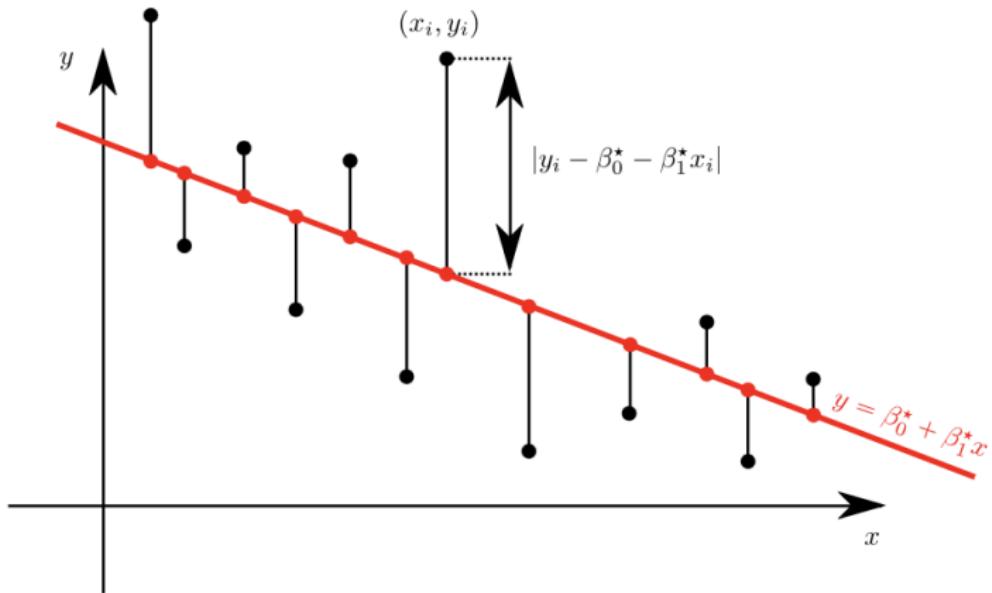
## Moindres carrés visualisation



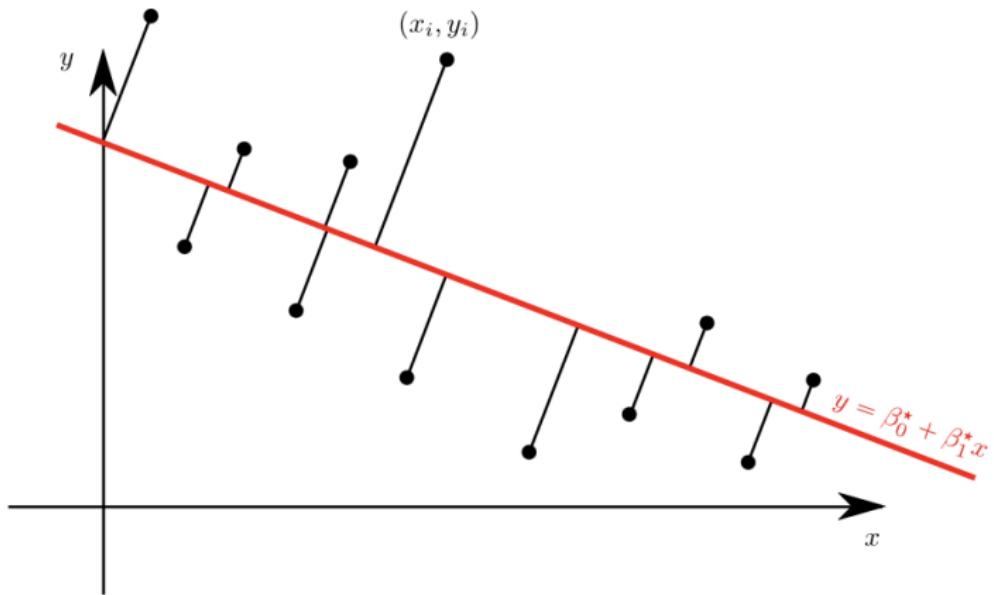
## Moindres carrés visualisation



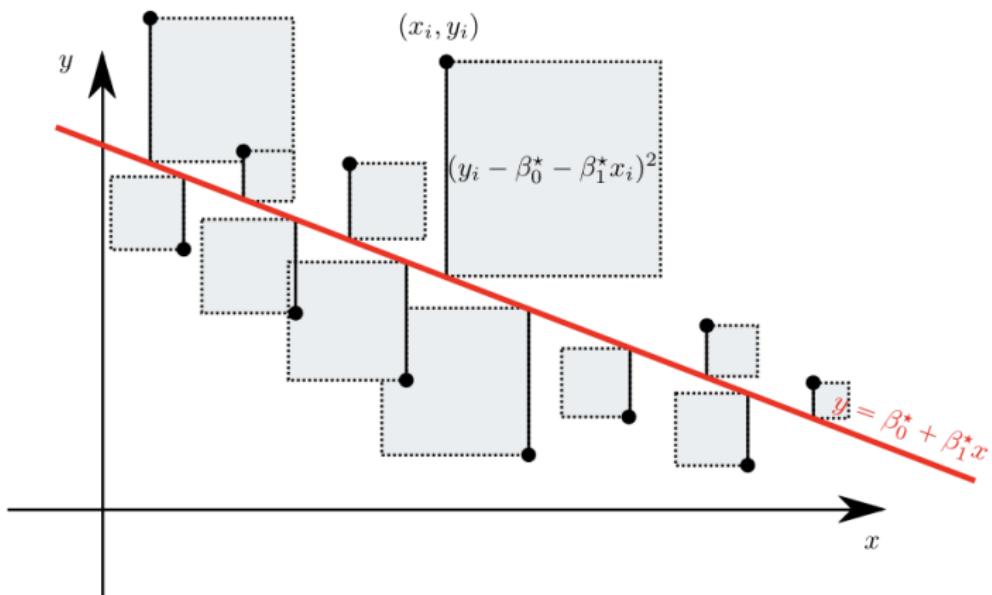
## Moindres carrés (totaux) visualisation



## Moindres carrés (totaux) visualisation



## Moindres carrés visualisation



## Estimation (*Learning*) par la méthode des moindres carrés

Pour des raisons mathématiques (e.g., simplicité computationnelle) on choisit de minimiser la somme des carrés des "erreurs".

### Définition

L'estimateur des moindres carrés est défini comme suit

$$(\hat{\beta}_0, \hat{\beta}_1) \in \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Autrement dit, on cherche le couple  $(\hat{\beta}_0, \hat{\beta}_1)$  qui réalise le minimum de la fonction de perte (*Loss function*)

$$L : (\beta_0, \beta_1) \mapsto \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

## Calcul de la solution

On cherche donc à minimiser  $(\beta_0, \beta_1) \mapsto \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$  en fonction de  $\beta_1$  et  $\beta_0$ .

La dérivée de  $\beta_0 \mapsto \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$  est donnée par

$$\sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)$$

La dérivée de  $\beta_1 \mapsto \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$  est donnée par

$$\sum_{i=1}^n 2x_i(y_i - \beta_0 - \beta_1 x_i)$$

Ce *L* s'appelle **fonction de perte** ou *fonction de coût*.

## Calcul de la solution

Ainsi,  $(\hat{\beta}_0, \hat{\beta}_1)$  est solution du système linéaire

$$\begin{cases} \sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

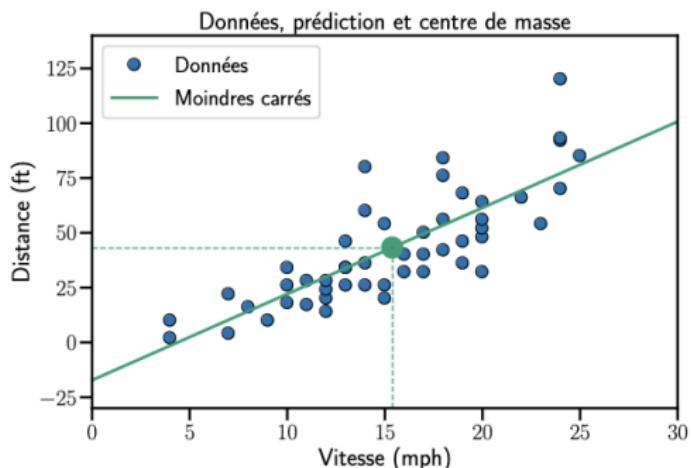
Soit

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \frac{1}{n} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - (\frac{1}{n} \sum_{i=1}^n x_i)^2} \\ \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i \end{cases}$$

En utilisant l'écriture,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  et  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ,

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - (\bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

## Interprétation



$$\text{vitesse} = 15.4 \text{ mph}$$

$$\text{distance} = 42.98 \text{ ft}$$

$\hat{\beta}_0 = -17.58 \text{ ft}$  (ordonnée à l'origine)

$\hat{\beta}_1 = 3.93 \text{ ft/mph}$  (pente de la droite)

Le centre de gravité du nuage de points (point en vert) est sur la droite de régression (estimée)

Prédiction : la prédition associée à l'observation moyenne  $\bar{x}$  est la moyenne  $\bar{y}$ .

## Prédictions et prédicteurs

### Définition

On appelle **prédicteur** une fonction qui à une nouvelle observation  $x_{n+1}$  associe une estimation de la variable à expliquer.

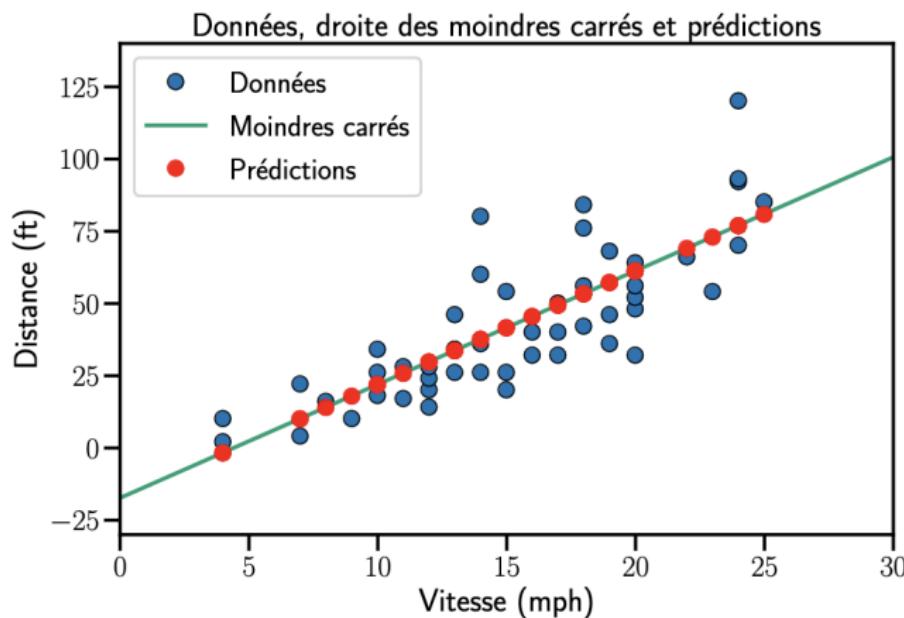
### Exemple

$$\text{pred}(x_{n+1}) = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

Une question importante sera d'évaluer la qualité d'un prédicteur.

## Prédictions

Les prédictions sont les points de la droite de régression.



## Validation croisée

En machine learning, la seule "vérité" connue est celle fournie par les observations disponibles.

En pratique, on découpe **aléatoirement** l'ensemble des données en deux sous parties :



- ▶ **ensemble d'apprentissage** ou d'entraînement (*train set*) : utilisé pour entraîner le modèle ie estimer les paramètres inconnus  
→  $\hat{\beta}_0, \hat{\beta}_1$
- ▶ **ensemble de validation** ou de test (*test set*) : utilisé pour mesurer les performances du modèle dans un contexte de prédiction.

## Validation croisée

L'objectif est d'estimer le risque d'erreur quand on utilise le modèle entraîné pour prédire la valeur associée à de nouveaux exemples.

- ▶ **ensemble d'apprentissage** ou d'entraînement (*train set*) : utilisé pour entraîner le modèle et estimer les paramètres inconnus dans le cas de la régression linéaire.  
→  $\hat{\beta}_0, \hat{\beta}_1$
- ▶ **ensemble de validation** ou de test (*test set*) : utilisé pour mesurer les performances du modèle dans un contexte de prédiction.  
→  $y_{\text{pred}} = \hat{\beta}_0 + x_{\text{test}} \hat{\beta}_1$   
On mesure alors le risque d'erreur à l'aide de l'**erreur en moyenne quadratique** (*mean square error*) :

$$\text{mse} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_{\text{pred}}(i) - y_{\text{test}}(i))^2$$

Plus faible est l'erreur, meilleur est le modèle : on obtient bien une mesure empirique<sup>2</sup> de la performance.

---

2. empirique : apprise à partir des observations

## Take home messages

- ▶ La régression linéaire est un ingrédient de base de l'intelligence artificielle.
- ▶ La régression linéaire simple permet de modéliser le lien entre une variable explicative et une variable cible, toutes deux définies sur  $\mathbb{R}$ .
- ▶ Les paramètres de la régression linéaire sont estimés en minimisant la fonction de perte des moindres carrés.
- ▶ On peut prédire la variable cible de nouvelles observations à l'aide de la régression linéaire (préalablement apprise).
- ▶ La validation du modèle est basée sur la validation croisée.
- ▶ On valide le modèle en calculant le score de l'erreur en moyenne quadratique.