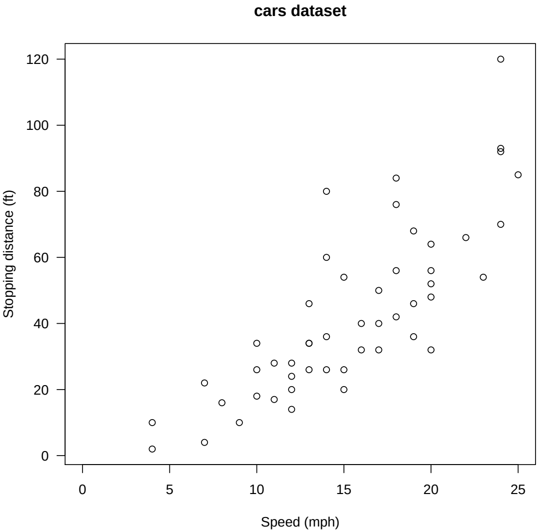
MDI220
Régression Linéaire

Joseph Salmon
Télécom ParisTech
http://josephsalmon.eu/

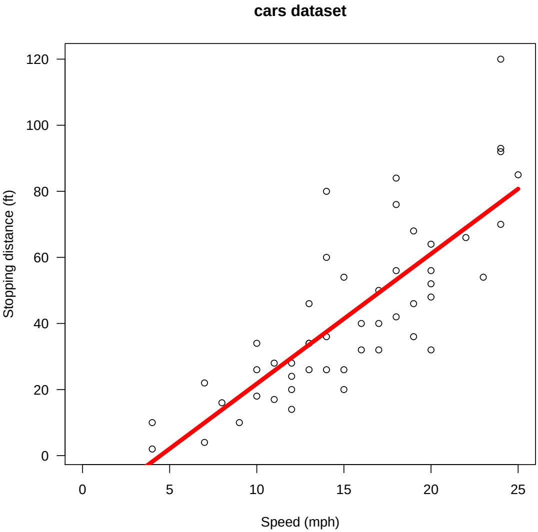
Point de départ en dimension deux

Exemple: distance de freinage en fonction de la vitesse des voitures pour 50 mesures



Point de départ en dimension deux

Exemple: distance de freinage en fonction de la vitesse des voitures pour 50 mesures



Commandes sous R:

```
attach(cars)
fm <- lm(dist ~ speed, data = cars)
```

De plus taper sous R la commande suivante:

```
fm$coefficients # $ sert pour les attributs
```

renvoit Intercept=-17.579095 et speed=3.932409 qui sont l'ordonnée à l'origine et la pente de la droite de la page d'avant.

Modélisation

Jeu d'observations: (y_i, x_i), pour $i = 1, \ldots, n$

Hypothèse de modéle linéaire ou de régression linéaire:

$$y \approx \theta_0 + \theta_1 x$$

avec θ_1 coefficient directeur et θ_0 ordonné à l'origine;
Rem: les deux paramètres sont inconnus

Exemple précédent

- $n = 50$
- y_i: temps de freinage de la voiture i
- x_i: vitesse de la voiture i
- l'hypothèse de la régression linéaire $\Leftrightarrow$ à postuler que le temps de freinage d'une voiture est proportionnel à sa vitesse

Modélisation (II)

On donne un sens au symbole $\approx$ de la manière suivante:

Modèle probabiliste

$$
\begin{aligned}
& y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \\
& \varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \ldots, n \\
& \mathbb{E}(\varepsilon) = 0
\end{aligned}
$$

où i.i.d. signifie indépendants et identiquement distribuées

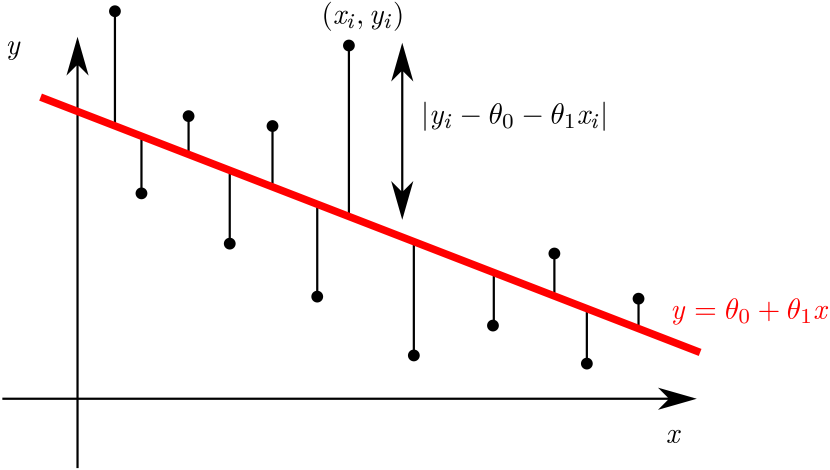
Interprétation

les $y_i - \theta_0 - \theta_1 x_i$, sont représentées par des variables aléatoires ε_i centrées (on parle aussi de **bruit blanc**) les erreurs entre le modéle théorique et les observations,
L'aspect aléatoire peut avoir diverses causes: bruit de mesures, variabilité dans une population, etc.

Objectif:

Estimer θ_0 et θ_1 par des quantités $\hat{\theta}_0$ et $\hat{\theta}_1$ dépendant des observations.

Estimateur des moindres carrés



Estimateur des moindres carrés (II)

Pour plusieurs raisons mathématiques on choisit de minimiser la somme des carrés des “erreurs” (plutôt que par exemple la somme des valeurs absolues des erreurs)

Formulation mathématique:

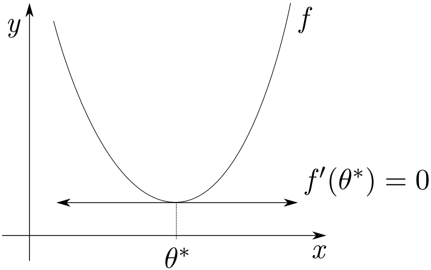
$$\boxed{(\hat{\theta}_0, \hat{\theta}_1) = \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2}$$

Optimisation dans $\mathbb{R}$

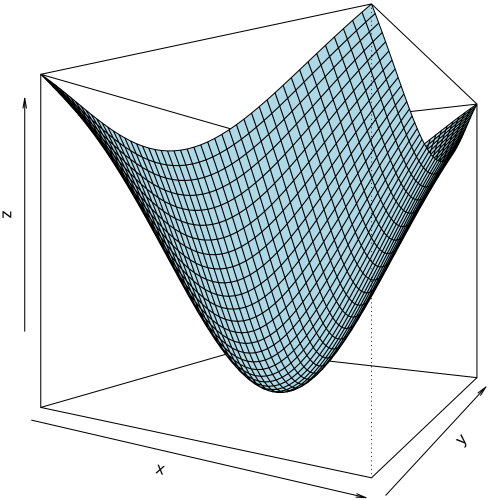
Théorème

Si une fonction $f : \mathbb{R} \to R$ est dérivable, alors un minimum $\theta^* \in \mathbb{R}$ de f doit vérifier la conditions nécessaire suivante, dite du premier ordre:

$$\boxed{f'(\theta^*) = 0}$$



Optimisation dans $\mathbb{R}^d$



Optimisation dans $\mathbb{R}^d$ (II)

Théorème

Si une fonction $f : \mathbb{R}^d \to R$ est différentiable ("régulière") un minimum $\theta^* \in \mathbb{R}^d$ de f doit vérifier la conditions nécessaire suivante, dite du premier ordre:

$$\boxed{\nabla f(\theta^*) = 0}$$

où $\nabla f(\theta) = (\frac{\partial f}{\partial x_1}(\theta), \ldots, \frac{\partial f}{\partial x_d}(\theta))$ est le gradient de f en θ^*

Rem: cette condition est nécessaire et suffisante si f est convexe

Retour aux moindres carrés

$$\boxed{(\hat{\theta}_0, \hat{\theta}_1) = \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2}$$

On cherche donc à minimiser une fonction de deux variables:

$$f(\theta_0, \theta_1) = f(\theta) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Condition du premier ordre:

$$\begin{cases} \frac{\partial f}{\partial \theta_0}(\hat{\theta}) = 2\sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) = 0 \\ \frac{\partial f}{\partial \theta_1}(\hat{\theta}) = 2\sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)x_i = 0 \end{cases}$$

Suite du calcul

Avec la notation $\bar{y}_n = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ et $\bar{x}_n = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

$$\begin{cases} \frac{\partial f}{\partial \theta_0}(\hat{\theta}) = 2\sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) = 0 \\ \frac{\partial f}{\partial \theta_1}(\hat{\theta}) = 2\sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)x_i = 0 \end{cases}$$

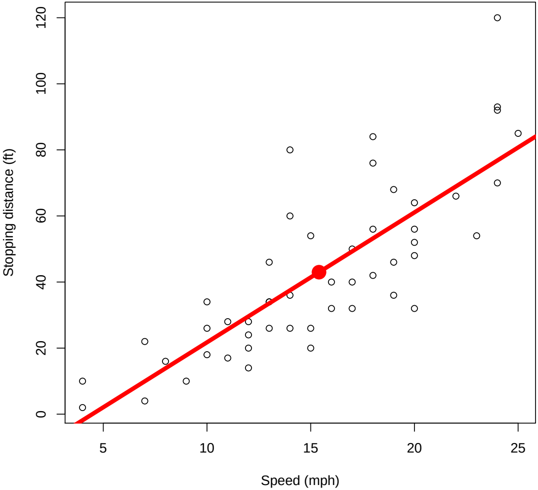
$$\Leftrightarrow$$

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

ATTENTION: formule **VRAIE** seulement si $\mathbf{x}$ est non constant
Preuve: EXO

Interprétation

Première équation: le point moyen appartient à la droite de régression estimée $(\bar{x}_n, \bar{y}_n) \in \{(x, y) \in \mathbb{R}^2 : y = \hat{\theta}_0 + \hat{\theta}_1 x\}$
Exemple: $\overline{speed} = 15.4; \overline{dist} = 42.98$



Interprétation (II)

Notation: $\mathbf{x} = (x_1, \ldots, x_n)^\top$ et $\mathbf{y} = (y_1, \ldots, y_n)^\top$

Deuxième équation:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

$$\Leftrightarrow$$

$$\hat{\theta}_1 = \mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \cdot \frac{\sqrt{\mathrm{var}_n(\mathbf{y})}}{\sqrt{\mathrm{var}_n(\mathbf{x})}}$$

où $$\mathrm{corr}_n(\mathbf{x}, \mathbf{y}) = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\mathrm{var}_n(\mathbf{x})}\sqrt{\mathrm{var}_n(\mathbf{y})}}$$

et $$\mathrm{var}_n(z) = \frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})^2$$

respectivement corrélations empiriques et variances empiriques

Exemple: $\mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \cdot \frac{\sqrt{\mathrm{var}_n(\mathbf{y})}}{\sqrt{\mathrm{var}_n(\mathbf{x})}} = 3.932409$

Recentrage

Nouveau model d'observation, dit recentré:

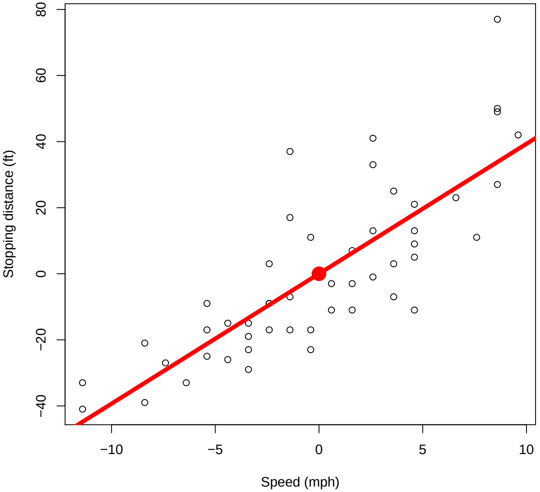
Si pour tout $i = 1, \dots, n$: $\begin{cases} x'_i = x_i - \bar{x}_n \\ y'_i = y_i - \bar{y}_n \end{cases} \Leftrightarrow \begin{cases} \mathbf{x}' = \mathbf{x} - \bar{x}_n \mathbf{1}_n \\ \mathbf{y}' = \mathbf{y} - \bar{y}_n \mathbf{1}_n \end{cases}$

si l'on note $\mathbf{1}_n = (1, \dots, 1)^\top$ et que l'on résout le programme des moindres carrés pour les $(\mathbf{x}', \mathbf{y}')$ alors

$$\begin{cases} \widehat{\theta'}_0 = 0 \\ \widehat{\theta'}_1 = \dfrac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x'_i y'_i}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x'^2_i} \end{cases}$$

Cela revient à définir l'origine comme le centre de gravité du nuage

Recentrage (II)



Recentrage + mise à l'échelle

Nouveau model d'observation, dit aussi **centré-réduit**:

$$\forall i = 1, \ldots, n : \begin{cases} x_i'' = (x_i - \bar{x}_n)/\sqrt{\mathrm{var}_n(\mathbf{x})} \\ y_i'' = (y_i - \bar{y}_n)/\sqrt{\mathrm{var}_n(\mathbf{y})} \end{cases} \quad \Leftrightarrow \begin{cases} \mathbf{x}'' = \dfrac{(\mathbf{x} - \bar{x}_n \mathbf{1}_n)}{\sqrt{\mathrm{var}_n(\mathbf{x})}} \\ \mathbf{y}'' = \dfrac{(\mathbf{y} - \bar{y}_n \mathbf{1}_n)}{\sqrt{\mathrm{var}_n(\mathbf{y})}} \end{cases}$$

En résolvant le programme des moindres carrés pour $(\mathbf{x}'', \mathbf{y}'')$ alors

$$\begin{cases} \widehat{\theta''}_0 = 0 \\ \widehat{\theta''}_1 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i'' y_i'' \end{cases}$$

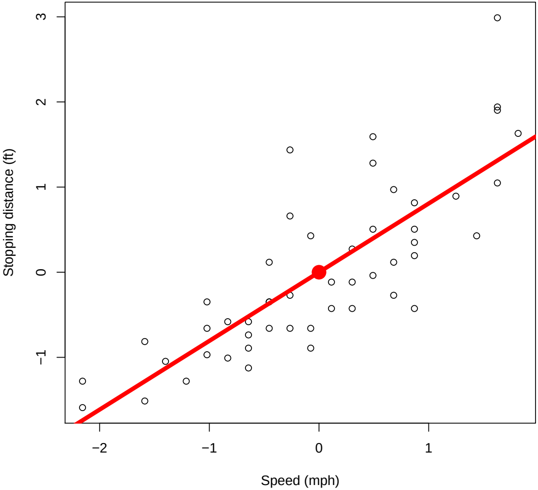
$\Leftrightarrow$ à définir le centre de gravité du nuage comme origine et à normaliser (pour la norme empirique $\| \cdot \|_n$) les vecteurs: $\mathbf{x}$ et $\mathbf{y}$

$$\|\mathbf{x}''\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i'')^2 = 1$$

$$\|\mathbf{y}''\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i'')^2 = 1$$

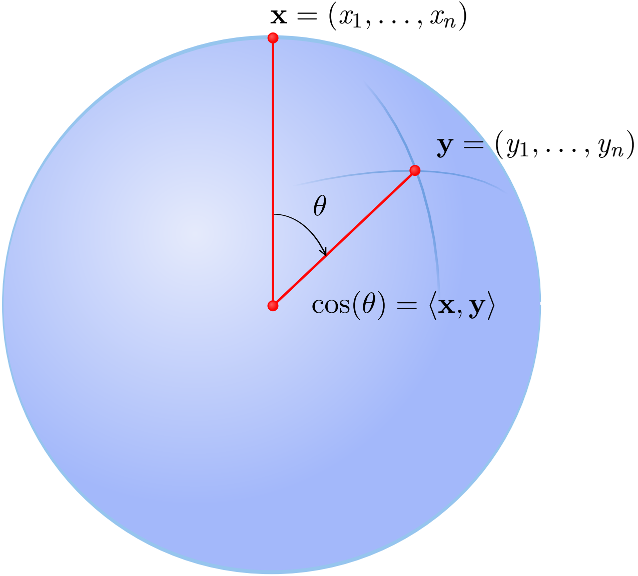
Recentrage + mise à l'échelle (II)

Interprétation: après recentrage / mise à l'échelle, on obtient:



Interprétation corrélation (cas centré-réduit)

Exemple dans le cas où $n = 3$ et $\|\mathbf{x}''\|_n^2 = \|\mathbf{y}''\|_n^2 = 1$



Définitions

Prédicteur

On appelle prédicteur la fonction qui à une nouvelle observation x_{n+1} propose une estimation: méthode

$$\mathrm{pred}(x_{n+1}) = \hat{\theta}_0 + \hat{\theta}_1 x_{n+1}$$

Résidus

On appelle résidus les différences entre les valeurs observées et la prédiction obtenues par notre méthode

$$r_i = y_i - \mathrm{pred}(x_i) = y_i - (\hat{\theta}_0 + \hat{\theta}_1 x_i)$$

Raison du choix des moindres carrés

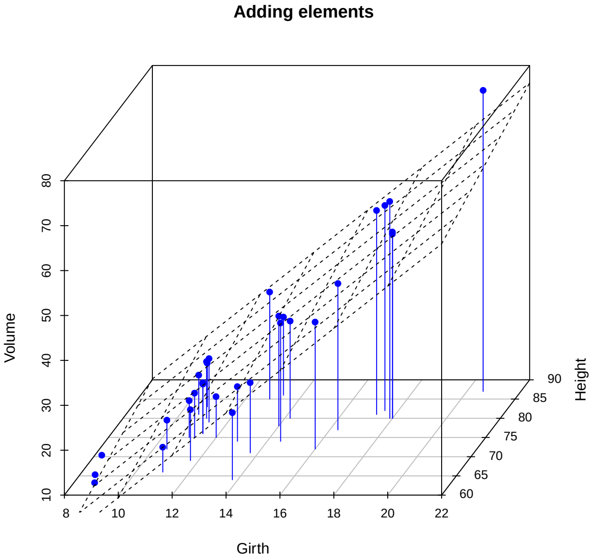
- Sous l'hypothèse que le bruit suit une loi gaussienne

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

  le maximum de (log)-vraisemblance amène à considérer les moindres carrés comme estimateur naturel de (θ_0, θ_1)
- Intérêt calculatoire: historiquement (invention par Gauss/Legendre fin XVIII ème début XIXeme siècle) il fallait trouve des formules fermées, *i.e.*, explicites.

Vers des modèles avec plus de deux variables

Volume d'arbres en fonction de leur hauteur / circonférence



Commandes sous R:

```
attach(tree)
fm <- lm(trees$Volume ~ trees$Girth + trees$Height)
```

De plus taper sous R la commande suivante:

```
require(scatterplot3d)
```

La commande

```
fm$coefficients
```

renvoie

`Intercept:-57.98 trees$Girth: 4.70 trees$Height: 0.33`

Modélisation

On dispose de p variables explicatives $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$

Modèle en dimension p

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{j,i} + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \ldots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

De manière équivalente:

$$\begin{cases} y_1 & = \theta_0 + \sum_{j=1}^{p} \theta_j x_{j,1} + \varepsilon_1 \\ & \vdots \\ y_n & = \theta_0 + \sum_{j=1}^{p} \theta_j x_{j,n} + \varepsilon_n \end{cases}$$

Dimension p

Modéle matricielle

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \; X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} \; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boxed{\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$$

Estimateur des moindres carrés

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \left(\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \right)$$

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left[y_i - \left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} \right) \right]^2$$

Noter: le fait que le signe n'est pas un signe d'égalité

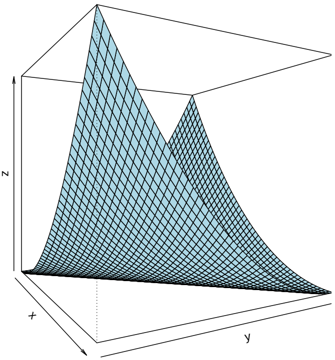
Formule fermée

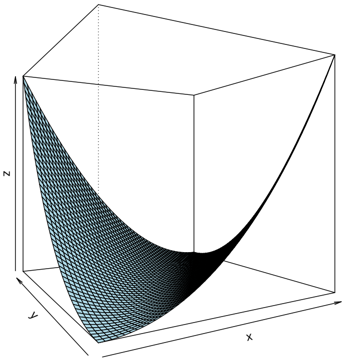
Si la matrice X est de plein rang (*i.e.*, si $X^\top X$ inversible)

$$\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Optimisation dans $\mathbb{R}^d$

Cas de fonction convexe (*e.g.*, $\theta \to \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$) dont le minimiseur n'est pas unique:





Site web pour aller plus

Wikipedia: Régression linéaire