Sciences des données

Un voyage initiatique

Cécile CAPPONI, Rémi EYRAUD, Hachem KADRI

LIS, Aix-Marseille Université, CNRS
Equipe QARMA

M1 Informatique

Plan

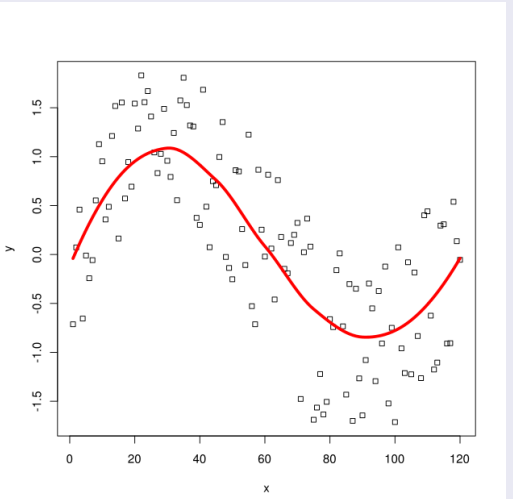
Problème de régression

Cas simple

On observe des données

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$$

On cherche à exprimer la dépendance entre le vecteur $\mathbf{x}$ et le réel y par une fonction, en se trompant le moins possible en généralisant



plusieurs classes de fonctions f, tq $f(\mathbf{x}) = y$

Calculer f à partir d'un échantillon $\{(\mathbf{x}_i, y)\}$, on veut calculer ce qui définit f : savoir ce qui la définit, donc savoir quoi chercher !

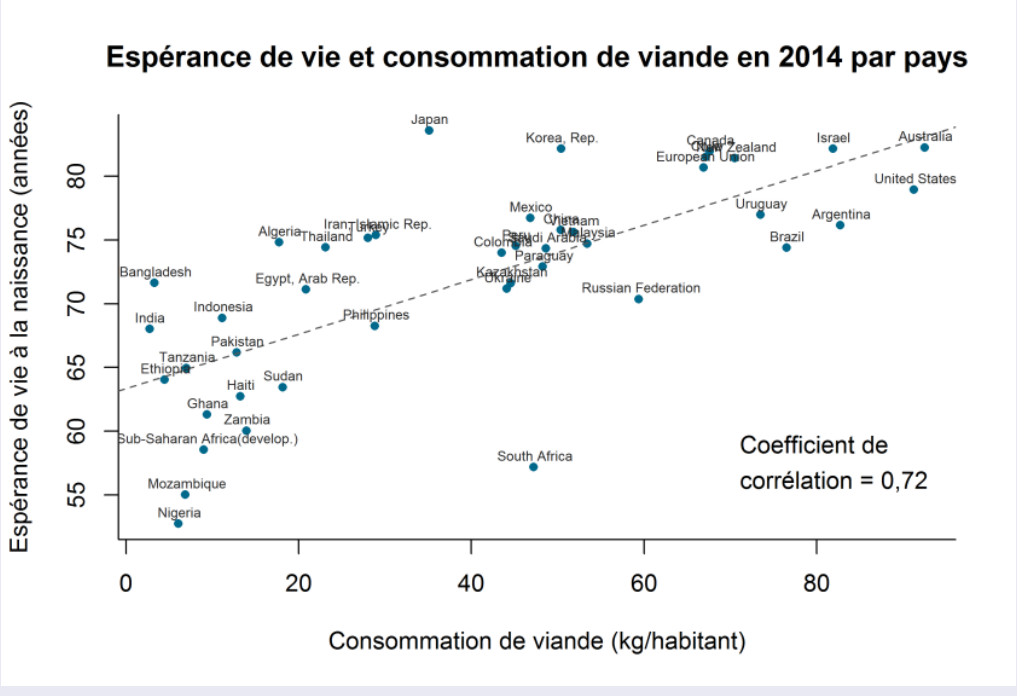
- La fonction est de la forme $f(\mathbf{x}) = a\mathbf{x} + b$: régression linéaire
- La fonction est un polynôme (ci-dessus) $f(\mathbf{x}) = a\mathbf{x}^3 + b\mathbf{x}^2 + c\mathbf{x} + d$
- Plus le polynôme est de degré élevé = plus on cherche de *paramètres* de la fonction = plus c'est compliqué
- Au delà des polynômes

Dans ce cours : régression linéaire

$$f(x) = ax + b$$

Souvenez vous...

Expliquer une variable par une (ou plusieurs) autre(s)



Un exemple : USCrime (L. Wasserman)

Crimes reliés à statistiques démographiques dans 47 états des USA : expliquer le taux de crimes par des variables mesurées.

1. R : Crime rate : # of offenses reported to police per million population
2. Age : The number of males of age 14-24 per 1000 population
3. S : Indicator variable for Southern states (0 = No, 1 = Yes) Ed : Mean # of years of schooling x 10 for persons of age 25 or older
4. Ex0 : 1960 per capita expenditure on police by state and local government
5. Ex1 : 1959 per capita expenditure on police by state and local government
6. LF : Labor force participation rate per 1000 civilian urban males age 14-24
7. M : The number of males per 1000 females
8. N : State population size in hundred thousands
9. NW : The number of non-whites per 1000 population
10. U1 : Unemployment rate of urban males per 1000 of age 14-24
11. U2 : Unemployment rate of urban males per 1000 of age 35-39
12. W : Median value of transferable goods and assets or family income in tens of $
13. X : The number of families per 1000 earning below 1/2 the median income

Un exemple : USCrime (suite)

R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
163.5	143	0	113	103	95	583	1012	13	102	96	36	557	194
57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
70.5	140	0	118	71	68	632	1029	7	15	100	24	526	174
167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
79.8	152	1	87	57	53	530	986	30	72	92	43	405	264
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Expliquer la variable R par les autres attributs

Calculer $R = f(\text{Age, S, Ed, etc., W,X})$

Bien calculer pour ne pas se tromper pour prédire les futurs crimes...

Comment calculer en faisant le moins d'erreurs possible ?

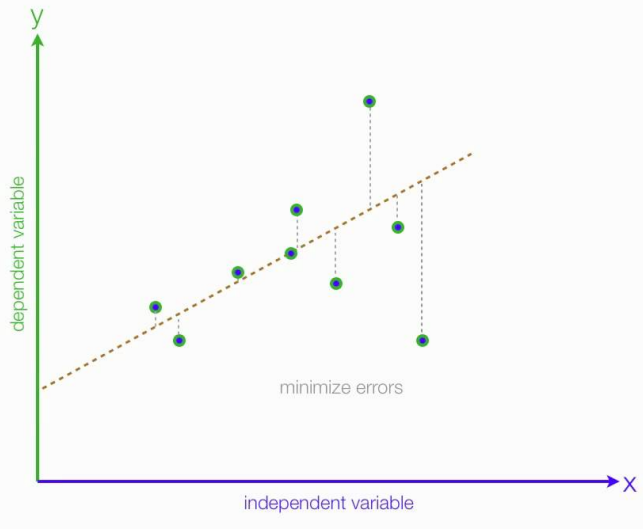
Comment calculer ?

Résoudre un système d'équations basé sur l'échantillon

- Que cherche t'on : une fonction, par exemple linéaire $y = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d + a_{d+1}$
- On cherche le vecteur de coefficients $a_1, \cdots a_{d+1}$ dans le cas d'une fonction linéaire

Comment faire le moins d'erreurs en généralisation ?

- Qu'est-ce qu'une erreur en régression ?
- Sur l'échantillon : minimiser l'écart de chaque point à la droite calculée (cela se mesure)
- Sur toute la population : cela s'estime



Modélisation de la régression

Un échantillon supposé i.i.d. selon une distribution jointe

- Une variable aléatoire $Z = (X, Y)$ à valeurs dans $\mathbb{R}^n \times \mathbb{R}$ (porteuse de liens entre les variables X – facteurs – et la variable Y à prédire – à expliquer)
- Les exemples sont des couples $(\mathbf{x}, y) \in \mathbb{R}^n \times \mathbb{R}$ tirés selon la distribution jointe $P(Z = (x, y)) = P(X = x)P(Y = y|X = x)$.
- Un échantillon S est un ensemble fini d'exemples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ i.i.d. selon P.

La fonction que l'on cherche et sa qualité

- Fonction de perte (loss) : $\ell(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$
- Fonction risque (erreur) : espérance mathématique de la fonction de perte

$$R(f) = \int \ell(y, f(\mathbf{x}))dP(\mathbf{x}, y) = \int_{\mathbb{R}^n \times \mathbb{R}} (y - f(\mathbf{x}))^2 dP(\mathbf{x}, y)$$

Modélisation de la régression (suite)

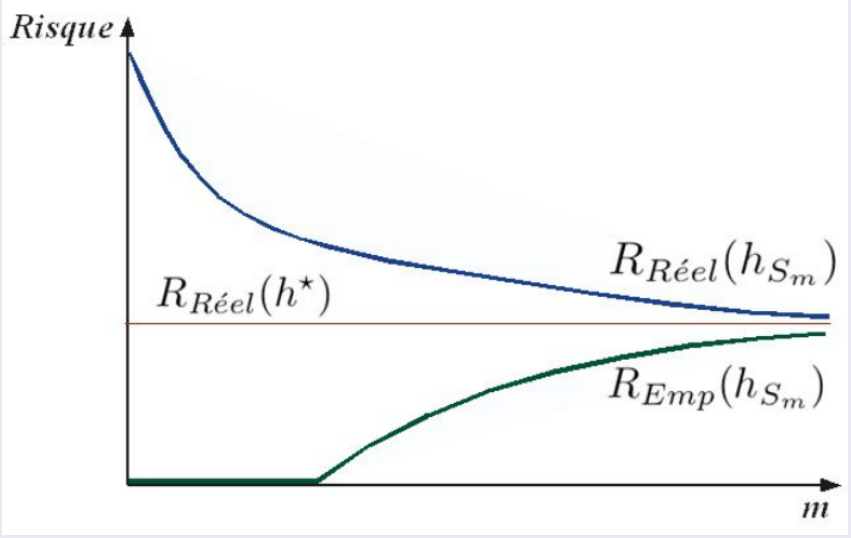
Le problème général de la régression

Etant donné un échantillon $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, trouver un classifieur f qui minimise le risque $R(f)$

La fonction de régression

- Bonne nouvelle : il existe une fonction qui minimise l'écart quadratique moyen : la *fonction de régression* $r(\mathbf{x}) = \int_Y y dP(y|\mathbf{x})$
- Mauvaise nouvelle : la fonction de régression est le plus souvent inaccessible (distribution inconnue, *cf* en classification)

Alternative : le principe MRE basé sur le risque empirique, celui qui se mesure sur l'échantillon d'apprentissage



Minimisation du risque empirique (MRE)

Risque empirique en régression

- Le risque empirique $R_{emp}(f)$ de f est la moyenne des carrés des erreurs de prédiction par f, calculée sur S

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$$

- Principe de minimisation du risque empirique : calculer

$$\underset{f}{\operatorname{argmin}} R_{emp}(f)$$

- Dans quel espace de fonctions chercher f ?
- Une notation parmi d'autres : $f_\theta(\mathbf{x})$ où θ symbolise les paramètres de la fonction à déterminer, par exemple :

$$f(\mathbf{x}) = a\sin^k(\mathbf{b}\dot{\mathbf{x}} + c), \theta = \{a, k, \mathbf{b}, c\}$$

$$f(\mathbf{x}) = \mathbf{a}\dot{\mathbf{x}} + b, \theta = \{\mathbf{a}, b\}$$

- Apprentissage = déterminer θ qui minimise le risque empirique

Régression linéaire

Définition de la régression linéaire

On suppose que

$$y = \langle \boldsymbol{\alpha}, \mathbf{x} \rangle + \beta + \epsilon$$

où

- $\mathbf{x}$ prend ses valeurs dans $\mathbb{R}^d$, la description des données
- $\boldsymbol{\alpha} \in \mathbb{R}^d$ et $\beta \in \mathbb{R}$ sont les paramètres θ à estimer
- ϵ est une variable aléatoire telle que $\mathbb{E}(\epsilon) = 0$ et $\mathbb{V}(\epsilon) = \sigma^2$ (variance indépendante de X) : du bruit
- $\langle \mathbf{u}, \mathbf{v} \rangle$ est le produit scalaire entre les vecteurs $\mathbf{u}$ et $\mathbf{v}$

La fonction de régression est

$$r(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{x} \rangle + \beta = \alpha_1 x_1 + \ldots \alpha_d x_d + \beta$$

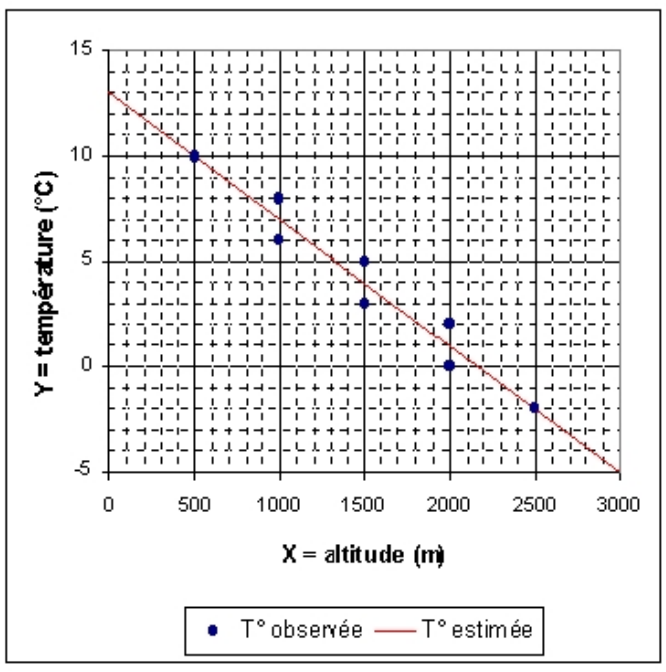
Régression linéaire - cas $d = 1$ (une seule variable facteur)

On cherche une équation de droite : deux paramètres !

On suppose que La fonction de régression est

$$r(x) = \alpha x + \beta$$

On suppose une corrélation linéaire entre x et y laquelle ?



Estimation de la fonction de régression à partir des données :

$$\hat{\alpha} = -\frac{1}{200} \text{ et } \hat{\beta} = 13$$

$$\hat{r}(x) = -\frac{1}{200}x + 13$$

Test : $\hat{r}(1800) = -9 + 13 = 4$

Régression linéaire : estimateurs des moindres carrés (MC)

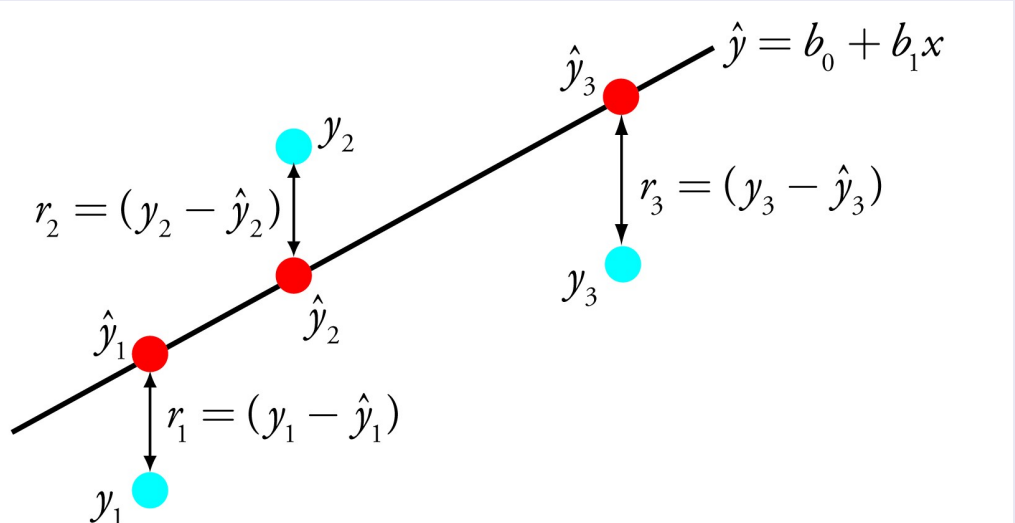
Récapitulatif et notations

En une dimension, on cherche α et β qui réalisent l'idéal $r(x) = \alpha x + \beta$ pour tout x de l'échantillon et en généralisation

- On ne trouvera pas l'idéal... Admettons.
- On va estimer α et β sur la base de l'échantillon : $\hat{\alpha}$ et $\hat{\beta}$ désignent les paramètres estimés via le principe MRE : sur la base de l'échantillon

Principe de l'estimateur des moindres carrés

Fondé sur l'écart quadratique de l'échantillon à la solution potentielle : à minimiser sur l'ensemble des points de l'échantillon



Régression linéaire par l'estimateur des moindres carrés, $d = 1$ dimension

L'estimateur des moindres carrés pour $d = 1$, formellement

Soit $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un échantillon de n exemples.
Les valeurs de $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent

$$\sum_{i=1}^{n} (y_i - (\hat{\alpha} x_i + \hat{\beta}))^2$$

sont

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\hat{\beta} = \overline{y} - \hat{\alpha} \overline{x}$$

où les moyennes considérées sont

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ et } \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Régression linéaire $d = 1$: la fonction de régression et son erreur apparente

Fonction de regression estimée

La fonction de régression estimée est alors

$$\hat{r}(x) = \hat{\alpha}x + \hat{\beta}.$$

Les erreurs estimées de cette fonction

- Erreur estimée sur un exemple = résidu

$$\widehat{\epsilon_i} = y_i - \widehat{y_i} = y_i - (\hat{\alpha}x_i + \hat{\beta})$$

- Variance estimée = erreur quadratique moyenne estimée (mse)

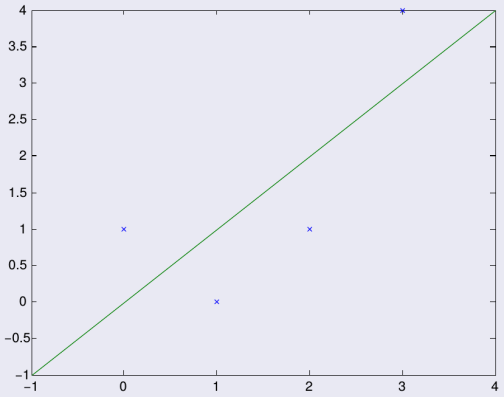
$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\epsilon_i}^2$$

Estimateurs des moindres carrés $d = 1$: exemple

Soit $S = \{(0,1),(1,0),(2,1),(3,4)\}$ notre échantillon

- Tracer l'échantillon sur un graphique
- Indiquer grossièrement une droite de régression linéaire
- Calculer la droite de régression linéaire par l'estimateur des moindres carrés, et son erreur

Solution



$$\overline{x} = 3/2, \overline{y} = 3/2, \hat{\alpha} = 1 \text{ et } \hat{\beta} = 0.$$

$$\widehat{\epsilon_1} = 1, \widehat{\epsilon_2} = -1, \widehat{\epsilon_3} = -1$$

$$\widehat{\epsilon_4} = 1 \text{ et } \hat{\sigma}^2 = 2$$

Simulateur sur internet : `https://bit.ly/2qwb5g2`

Propriétés de l'estimateur des moindres carrés

Estimateur sans biais

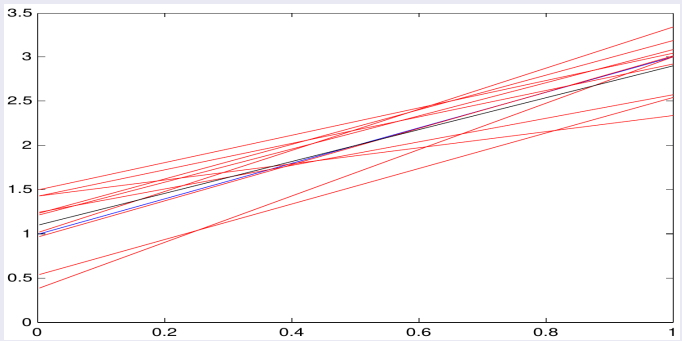
$\hat{\alpha}, \hat{\beta}$ et $\hat{\sigma}^2$ sont des *estimateurs non biaisés* de α, β et σ^2

- Répéter *m* fois l'apprentissage des paramètres avec *m* échantillons différents de la même population
- La moyenne des *m* paramètres estimés converge vers les paramètres du modèle

Illustration (*m* = 10 expériences)

- X prend 11 valeurs équidistantes dans $[0, 1]$; $Y = 2 * X + 1 + Norm(0, 1)$
- Les *m* échantillons diffèrent en raison du bruit dans l'étiquette

$S_1 = \{(0, 1.24), (0.1, 0.99), (0.2, 1.31), \cdots, (1, 2.95)\}$,
$S_2 = \{(0, 0.88), (0.1, 0.99), (0.2, 1.55), \cdots, (1, 3.17)\}$, etc.



- en bleu : la droite de régression
- en rouge : chaque estimation
- en noir : la moyenne des estimations.

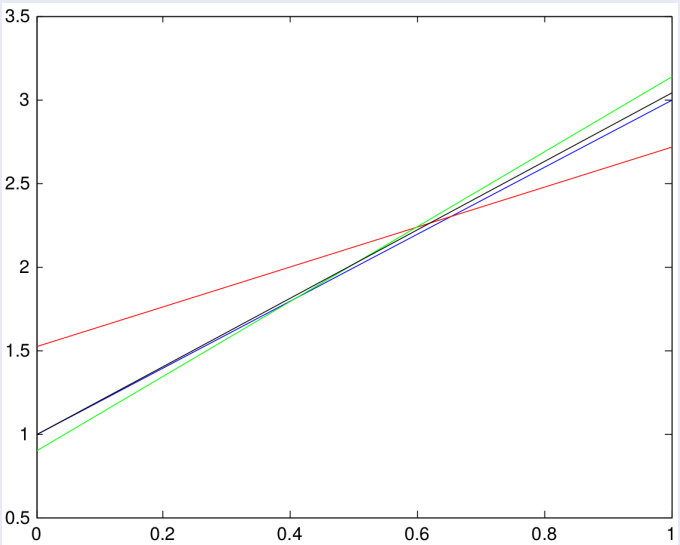
Propriétés de l'estimateur des moindres carrés (suite)

Estimateur consistant (convergent)

plus on dispose d'observations (n grand), plus les estimations se rapprochent des paramètres du modèle.

Illustration

- X prend n valeurs équidistantes dans $[0, 1]$; $Y = 2 * X + 1 + Norm(0, 1)$.
- Faisons varier n, de 11 à 1001



- en bleu : la droite de régression
- en rouge : $n = 11$
- en vert : $n = 101$
- en noir : $n = 1001$.

Régression linéaire multivariée : $d > 1$

Plusieurs variables facteurs, une variable expliquée

Soit $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ l'échantillon d'apprentissage, $\mathbf{x}_i \in \mathbb{R}^d$, $y \in \mathbb{R}$

- Soit X la matrice $n \times (d+1)$ dont la i-ème ligne est $\mathbf{x}_i$, 1 : $i$ème exemple, composé de $d+1$ valeurs (la dernière indique l'écart à l'origine)
- Soit Y le vecteur colonne composé des étiquettes y_i.

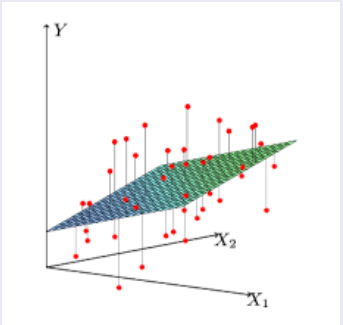
Estimateur des moindres carrés, $d > 1$

Les paramètres θ du modèle sont : $\boldsymbol{\alpha}$ un vecteur de $\mathbb{R}^d$, β un réel
La fonction de régression est $r_\theta(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\alpha} \rangle + \beta$

- L'estimateur des moindres carrés est

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\beta} \end{pmatrix} = (X^T X)^{-1} X^T Y$$

où X^T = matrice transposée de X



- Si $X^T X$ n'est pas inversible, ou si $det(X^T X) \simeq 0$, . . . il est nécessaire de transformer le problème (second semestre)

Exemple de régression linéaire multivariée

$S = \{((0,0),-1),((0,1),1),((1,0),1),((1,1),1)\}$.
On a

$$X = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, X^T = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \text{ et } Y = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

On vérifie que

$$X^T X = \begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 2 & 4 \end{pmatrix}, (X^T X)^{-1} = \begin{pmatrix} 1 & 0 & -1/2 \\ 0 & 1 & -1/2 \\ -1/2 & -1/2 & 3/4 \end{pmatrix}$$

$$(X^T X)^{-1} X^T = \begin{pmatrix} -1/2 & -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & 1/2 \\ 3/4 & 1/4 & 1/4 & -1/4 \end{pmatrix} \text{ et } (X^T X)^{-1} X^T Y = \begin{pmatrix} 1 \\ 1 \\ -1/2 \end{pmatrix}$$

soit

$$\hat{\alpha} = (1,1) \text{ et } \hat{\beta} = -1/2.$$

Prédiction $r((0.2,0.7) = 0.2\alpha_1 + 0.7\alpha_2 + \beta = 0.2 + 0.7 - 0.5 = 0.4$