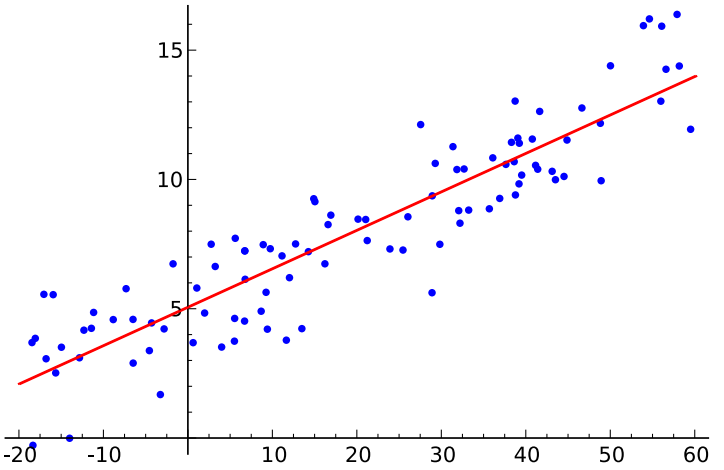
INFO601_CMI Algorithmique numérique

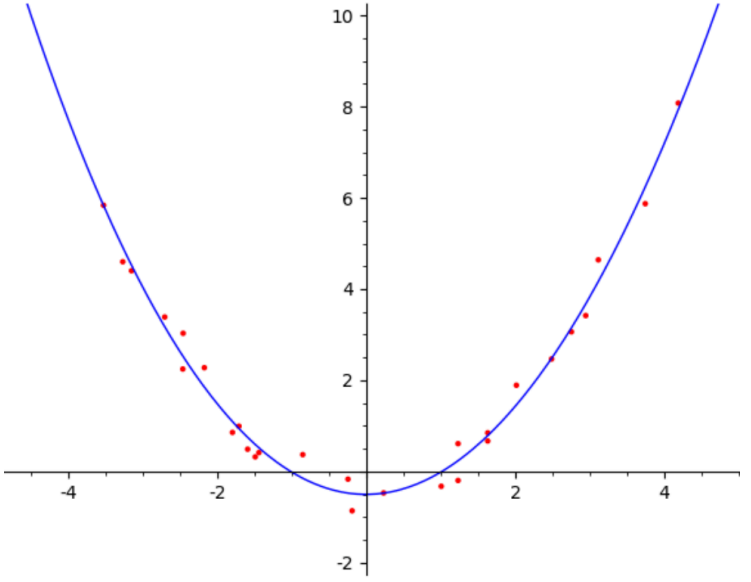
Introduction au domaine via la méthode des moindres carrés

[Jacques-Olivier Lachaud]

Modèle prédictif à partir de données

- *données observées* $[(x_1, y_1), ..., (x_n, y_n)]$
- peuvent conduire à une *prédiction* ?
- *prédiction* : $f(x)$ avec $\forall i, f(x_i) \approx y_i$
- famille paramétrée de modèles $f_{\boldsymbol{\alpha}}$
- modèle prédictif le plus "*pertinent*" ?
  $\Rightarrow$ Trouver $\boldsymbol{\alpha}^*$ tel que $f_{\boldsymbol{\alpha}^*}$ prédit au mieux les y à partir des x





modélisation

- identifier une famille de fonction $f_{\boldsymbol{\alpha}}$
- caractériser les résultats pertinents

optimisation

- écrire le problème sans forme soluble
- chercher le meilleur paramètre $\boldsymbol{\alpha}^*$

Méthode des moindres carrés ordinaires

[Gauss, Legendre, circa 1800]

modélisation

- forme le vecteur des erreurs ou résidu

$$\boldsymbol{r} = \begin{pmatrix} y_1 - f_{\boldsymbol{\alpha}}(x_1) \\ \vdots \\ y_n - f_{\boldsymbol{\alpha}}(x_n) \end{pmatrix}$$

- erreur quadratique (norme euclidienne au carré)

$$E(\boldsymbol{\alpha}) = \|\boldsymbol{r}\|^2 = \sum_{i=1}^{n} \left(y_i - f_{\boldsymbol{\alpha}}(x_i)\right)^2$$

- $E(\boldsymbol{\alpha})$ est une fonction coût
- modèle pertinent = petit coût

optimisation

- $\boldsymbol{\alpha}^* := \arg\min_{\boldsymbol{\alpha} \in D} E(\boldsymbol{\alpha})$
- problème **difficile** en général
- min $E \Rightarrow$ dérivées nulles

$\Rightarrow$ gradient $\boldsymbol{\nabla} E = 0$

- trouver les zéros du gradient
- algorithmes de descente en gradient
- résolution de systèmes linéaires

Régression linéaire simple

- ajustement droite $f_{\boldsymbol{\alpha}}(x) = \alpha_1 x + \alpha_2$
- erreur quadratique $E(\ \alpha_1\ ,\ \alpha_2\ ) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \ \alpha_1\ x_i + \ \alpha_2\ )^2$

$$E(\alpha_1, \alpha_2) = (\alpha_1 \ \ \alpha_2) \underbrace{\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & \sum 1 \end{pmatrix}}_{M} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} - 2(\alpha_1 \ \ \alpha_2) \underbrace{\begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}}_{\boldsymbol{v}} + \underbrace{\sum y_i^2}_{c}$$

$$= \boldsymbol{\alpha}^\mathsf{T} M \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{v} + c$$

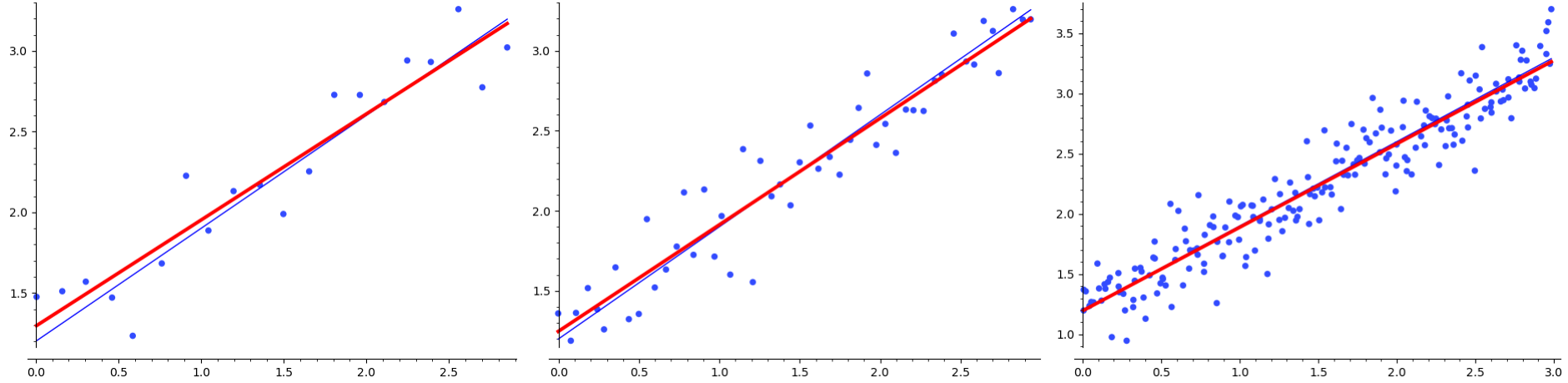
- *forme quadratique définie positive* $\Rightarrow E$ a un seul minimum
- **optimisation** mieux ici de chercher le zéro du gradient de E

$$\boldsymbol{\nabla} E = \begin{pmatrix} \frac{\partial E}{\partial \alpha_1} \\ \frac{\partial E}{\partial \alpha_2} \end{pmatrix} = \begin{pmatrix} 2m_{11}\alpha_1 + (m_{12} + m_{21})\alpha_2 - 2v_1 \\ (m_{12} + m_{21})\alpha_1 + 2m_{22}\alpha_2 2v_2 \end{pmatrix} = \underbrace{(M + M^\mathsf{T})}_{2M}\boldsymbol{\alpha} - 2\boldsymbol{v}$$

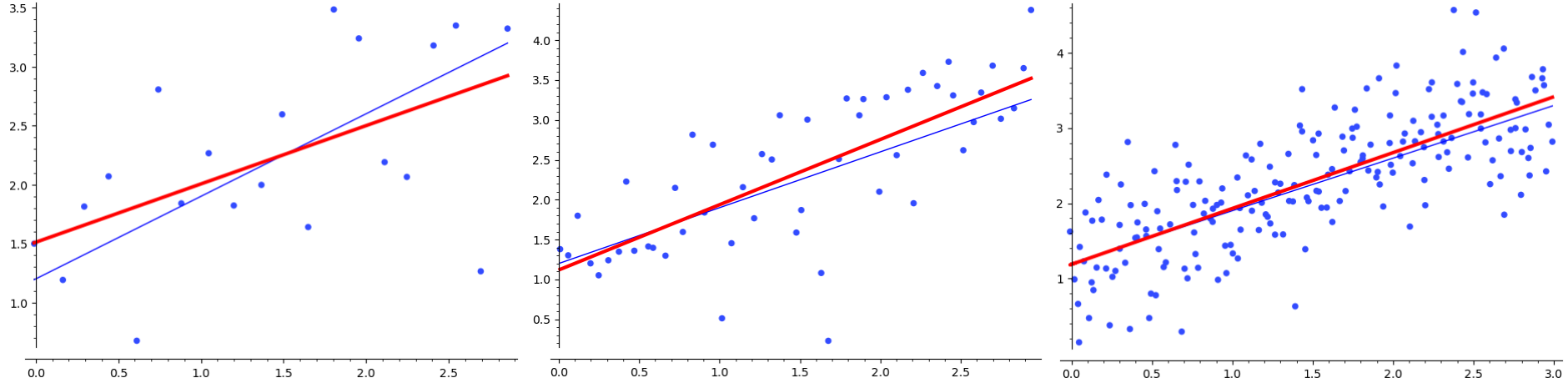
$\boldsymbol{\alpha}^*$ solution de $\boldsymbol{\nabla} E = 0 \Leftrightarrow M\boldsymbol{\alpha} = \boldsymbol{v}$, i.e. $\boldsymbol{\alpha}^* = M^{-1}\boldsymbol{v}$

Régression linéaire simple

perturbation $\sigma = 0.2$



perturbation $\sigma = 0.6$



Des garanties théoriques

Si $y_i = \alpha_1 x_i + \alpha_2 + \varepsilon_i$, avec ε_i perturbation de la i-ème donnée

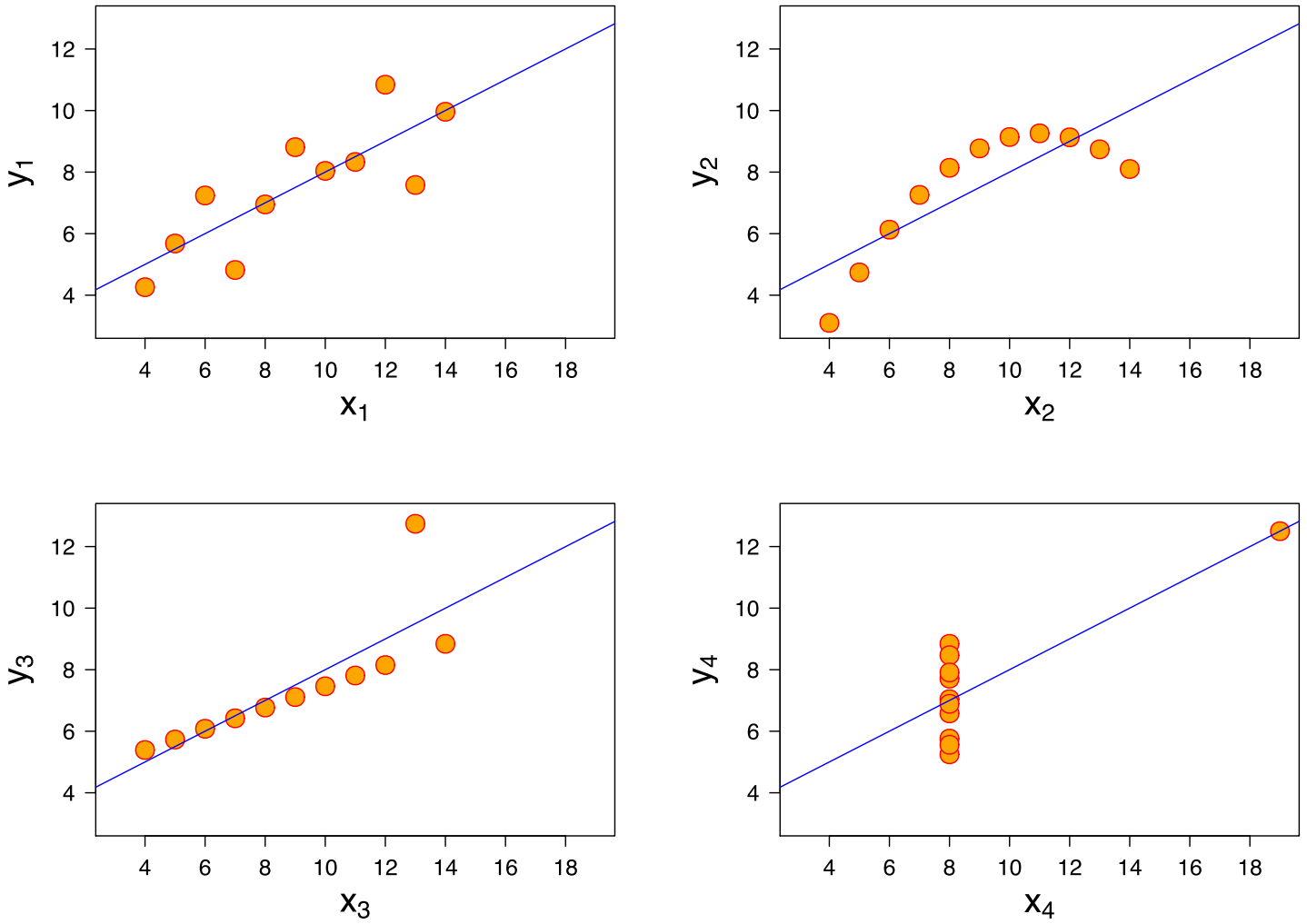
$$\underbrace{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}}_{\text{écriture matricielle}} \Leftrightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Thm [Gauss-Markov]: parmi tous les estimateurs linéaires non biaisés, l'estimateur par moindres carrés présente une variance minimale, i.e. $\|\alpha^* - \alpha\|^2$ est minimal. C'est le **BLUE** !

hypothèses

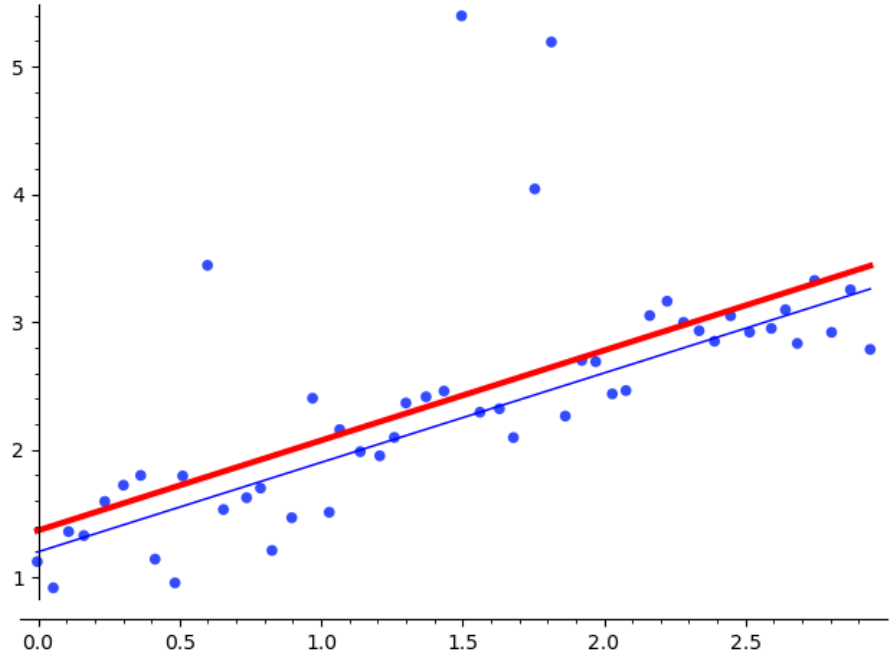
- $\mathbb{E}(\varepsilon_i) = 0$ "les erreurs sont sans biais"
- $\mathbb{V}\text{ar}(\varepsilon_i) = \sigma^2 < +\infty$ "les erreurs ont même variance, finie"
- $\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$ "les erreurs sont non corrélées"
- *estimateur linéaire* de α_j: $\hat{\alpha}_j = c_1 y_1 + ... + c_n y_n$, avec c_i qui peuvent dépendre des x_i

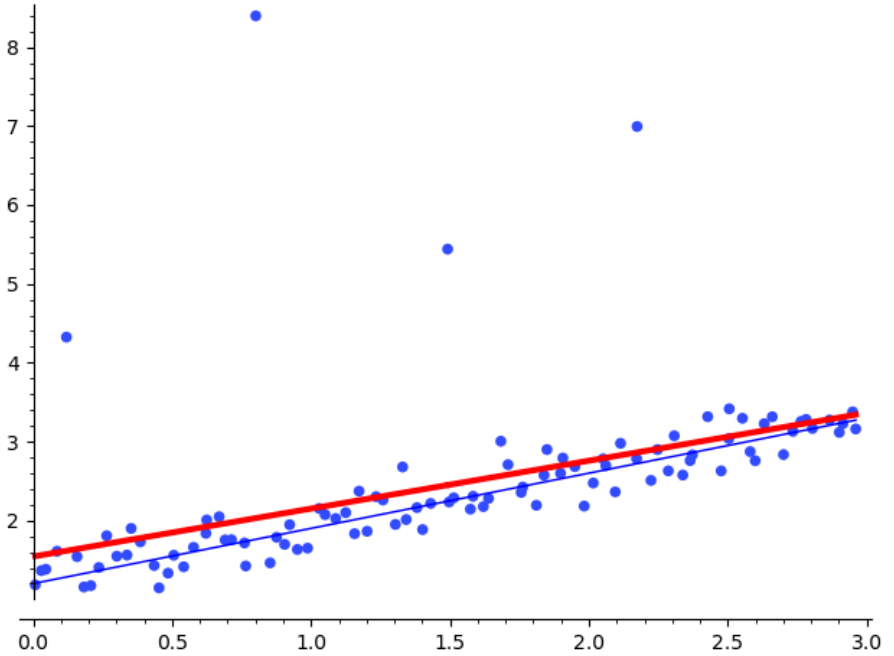
Attention à l'interprétation du résultat !

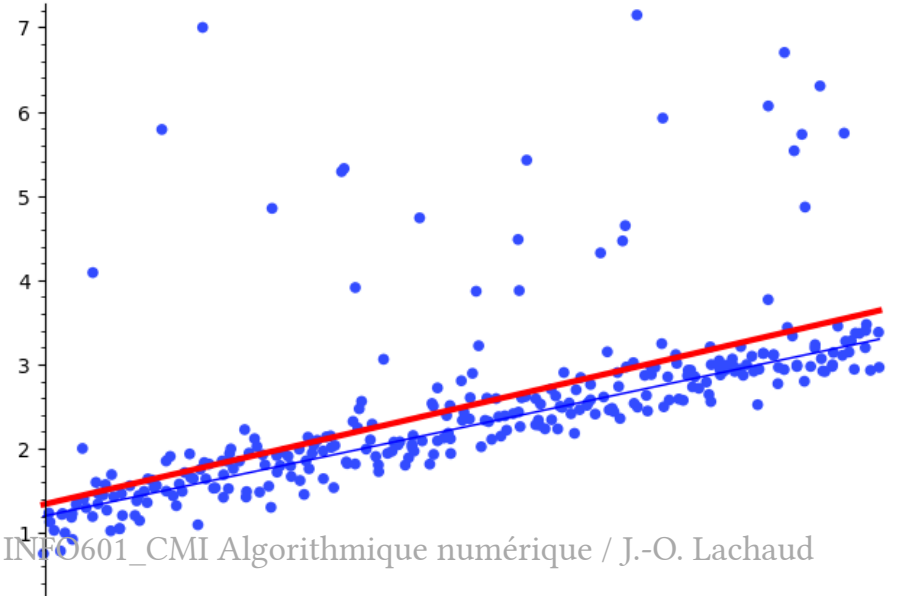


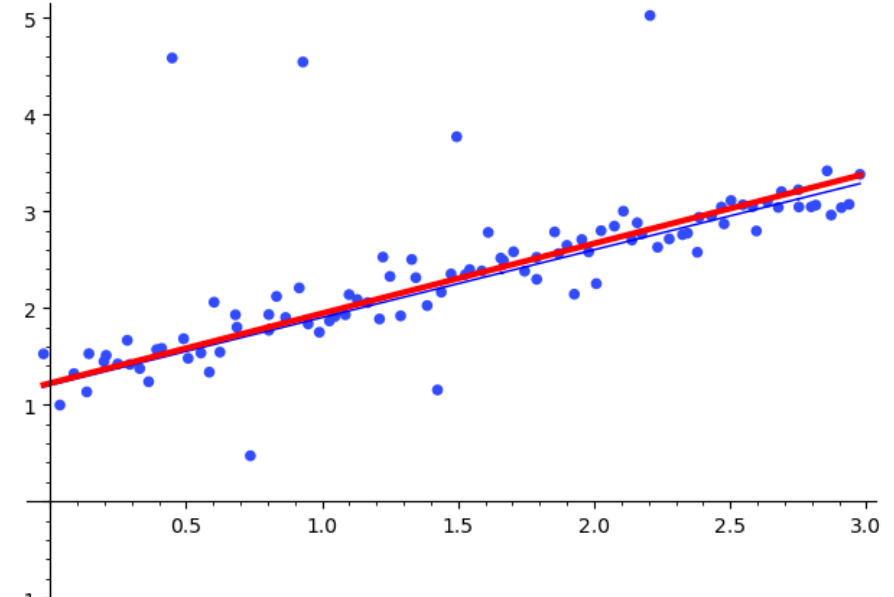
[Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21]

Problème des données aberrantes (outliers)









Elimination des données aberrantes

utilisation des résidus

- enlever 10% des données ayant le plus gros résidu

heuristiques

- random forest, isolation forest, distance au barycentre des k-NN

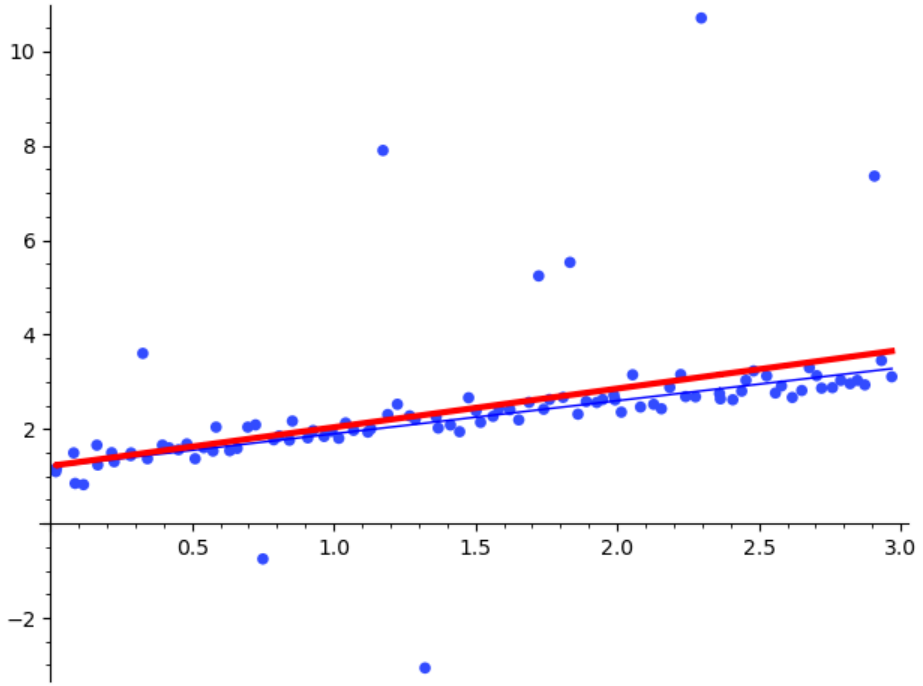
Autres méthodes statistiques

- Theil–Sen estimator: pente α_1 en prenant le médian des pentes de tous les couples de points.

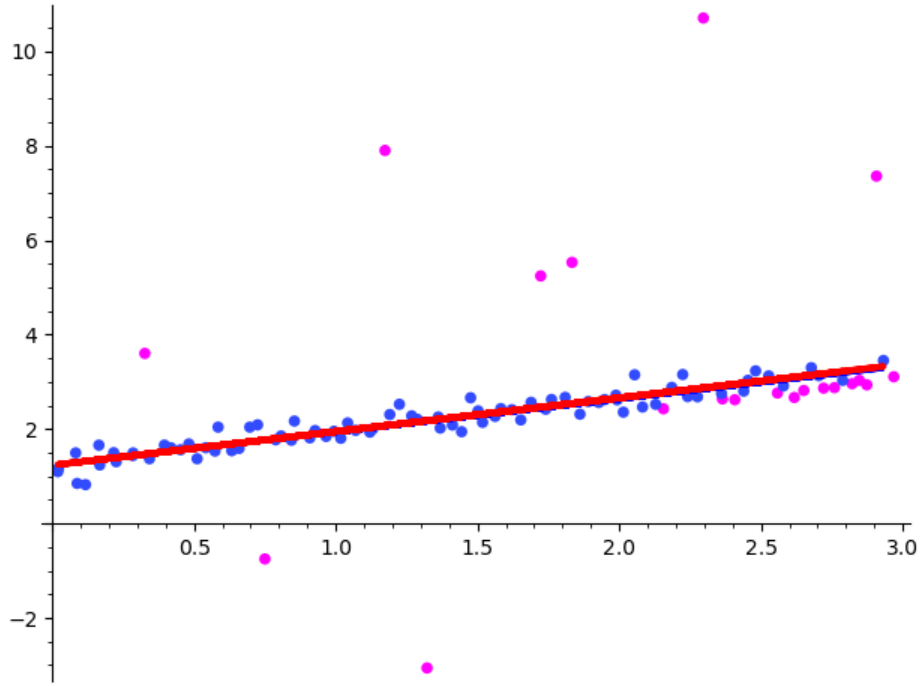
Problème de paramétrages, de choix de modèles, d'hypothèses sur les données, de temps de calcul, ...

Exercice : Complexité de la régression linéaire (n données) ? De l'estimateur Theil-Sen ?

Elimination des données aberrantes



sans filtrage

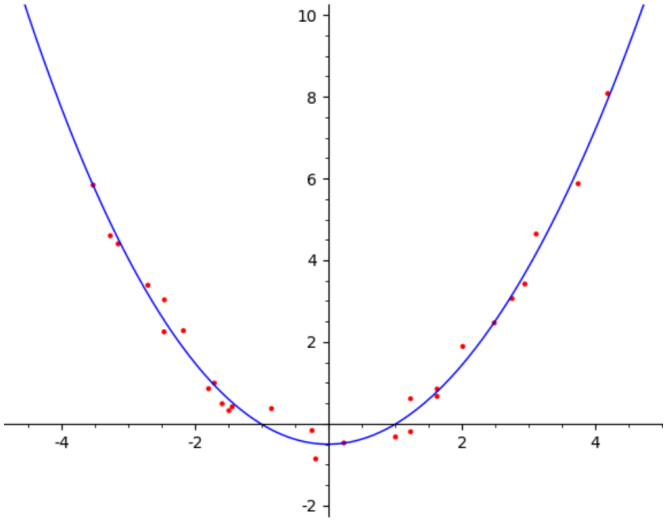


enlève 20% des données + recalcul

- on calcule α^* pour toutes les données
- on trie les données selon leur résidus r^2
- on garde les premiers 80% des données
- on recalcule α^* pour ces données

Et si on cherche des modèles plus complexes que la droite ?!

Régression linéaire multiple



- on suppose que les données $(x_i, y_i)_{i=1,\dots,n}$ suivent une loi parabolique
- pour chaque donnée, $y_i = \alpha_1 x_i^2 + \alpha_2 x_i + \alpha_3 + \varepsilon_i$, avec ε_i une erreur

$$\underbrace{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}}_{\text{écriture matricielle}} \Leftrightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La relation est encore linéaire !

Régression linéaire multiple

- vecteur résidu : $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}$, erreur quadratique (ou fonction coût):

$$E(\boldsymbol{\alpha}) = \|\boldsymbol{r}\|^2 = \boldsymbol{r}^\mathsf{T}\boldsymbol{r} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^\mathsf{T}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})$$
$$= \boldsymbol{\alpha}^\mathsf{T}\underbrace{\boldsymbol{X}^\mathsf{T}\boldsymbol{X}}_{M}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\mathsf{T}\underbrace{\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}}_{\boldsymbol{v}} + \underbrace{\boldsymbol{Y}^\mathsf{T}\boldsymbol{Y}}_{c}$$

Rappel: régression linéaire simple

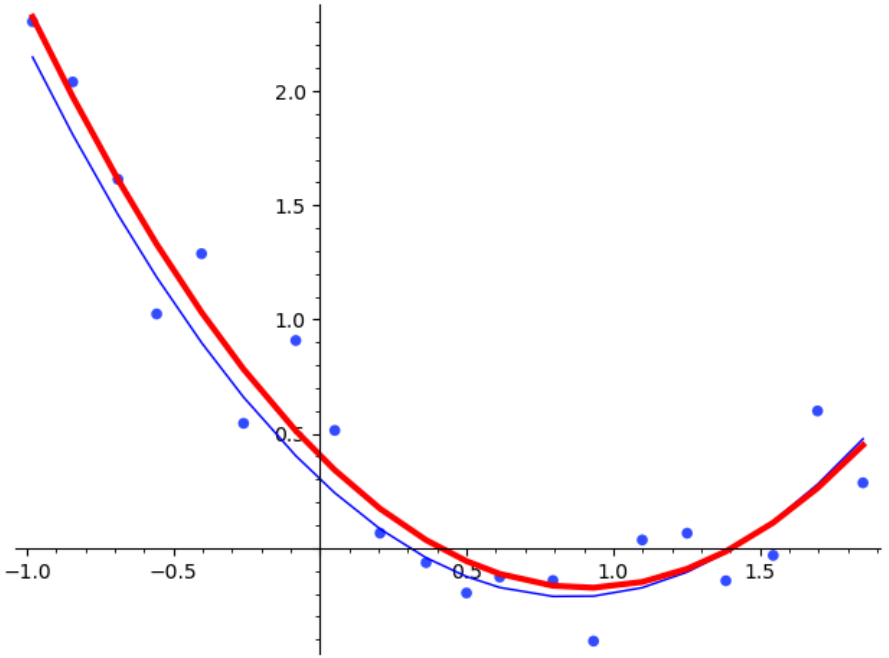
$$E(\alpha_1, \alpha_2) = (\alpha_1 \;\; \alpha_2)\underbrace{\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & \sum 1 \end{pmatrix}}_{M}\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix} - 2(\alpha_1 \;\; \alpha_2)\underbrace{\begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}}_{\boldsymbol{v}} + \underbrace{\sum y_i^2}_{c}$$

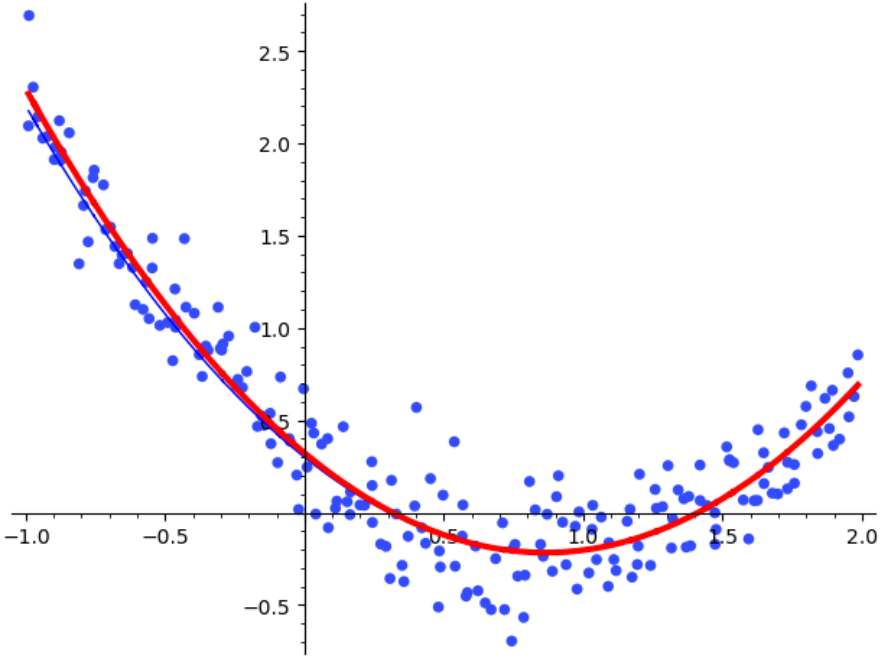
- **optimisation** en cherchant le zéro du gradient de E

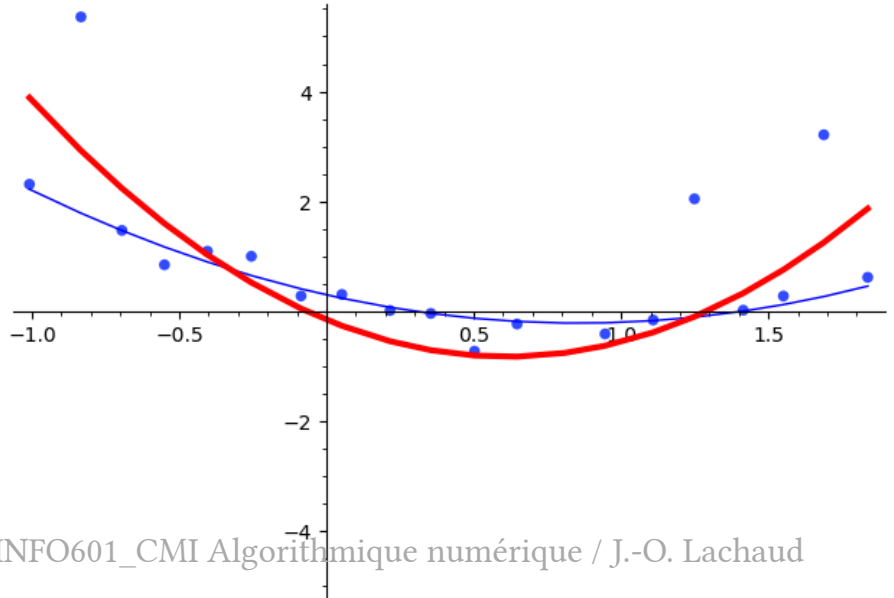
$$\boldsymbol{\nabla} E = \frac{\partial E}{\partial \boldsymbol{\alpha}} = 2\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\alpha} - 2\boldsymbol{X}^\mathsf{T}\boldsymbol{Y} = 2M\boldsymbol{\alpha} - 2\boldsymbol{v}$$

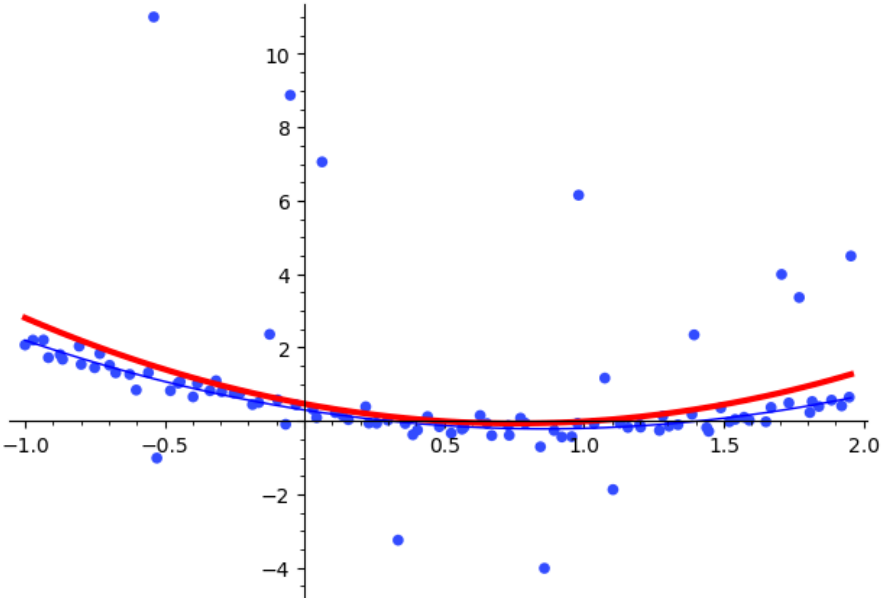
$$\boldsymbol{\alpha}^* \text{ solution de } \boldsymbol{\nabla} E = 0 \Leftrightarrow \boldsymbol{\alpha}^* = M^{-1}\boldsymbol{v} = \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}$$

Régression linéaire multiple









Méthode des moindres carrés

Régression linéaire multiple

- ajuster un polynôme de degré quelconque reste linéaire
- ajuster un plan à des données (x_i, y_i, z_i) reste linéaire
- ajuster une surface polynomiale à des données (x_i, y_i, z_i) reste linéaire

Moindres carrés généralisés

- donner un poids $w_i = \frac{1}{\sigma_i}$ différent à chaque donnée (typique d'une incertitude sur une mesure)
- donner une matrice de variance/covariance Σ entre les données

$\Rightarrow$ une approche similaire à la précédente fonctionne

Moindres carrés non linéaires

- ajuster une fonction $y = \cos(\alpha_1 x + \alpha_2)$ n'est pas linéaire

$\Rightarrow$ mais on peut ajuster $\text{acos}(y) = \alpha_1 x + \alpha_2$!

- ajuster une fonction $y = \exp\left(-\alpha_1 (x - \alpha_2)^2\right)$ n'est pas linéaire

Quelques exercices

Exercice 1: On cherche à ajuster un polynôme de degré 3 à des données $(x_i, y_i)_{i=1,\dots,n}$. Explicitez la relation matricielle $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}$.

Exercice 2: On vous donne les données suivantes : $(-1, 2), (1, 3), (3, 5), (4, 6), (6, 8)$ Trouvez la droite de meilleur ajustement.

Exercice 3: On dispose d'un nuage de points $(x_i, y_i, z_i)_{i=1,\dots,n}$. Proposez une approche pour trouver un plan approchant ces points.

Exercice 4: Est-ce que l'on obtient la même droite d'ajustement si on ajuste plutôt les x par rapport aux y ? On suppose donc que les données suivent

$$x_i = \beta_0 y_i + \beta_1 + \varepsilon'_i$$