

Diagnostic de la Régression

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

20 avril 2026

Diagnostic de la régression

Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

La divergence d'un individu

Les variables sont elles toutes pertinentes

Rappels

- ▶ Les données : (\mathbf{x}_i, y_i) (n observations)

- ▶ Le Modèle ($p + 1$ inconnues)

$$\mathbf{y} = X\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad y = a_0 + a_1x_1 + \dots + a_px_p + \varepsilon$$

- ▶ Le principe de projection

$$X^\top \boldsymbol{\varepsilon} = 0 \quad \Leftrightarrow \quad X^\top (\mathbf{y} - X\boldsymbol{\alpha}^*) = 0$$

- ▶ Les coefficients de la régression

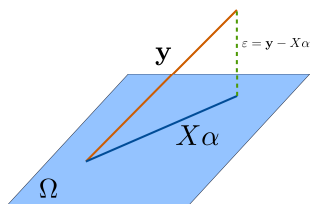
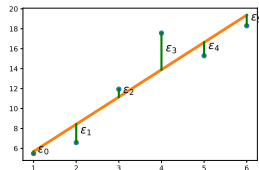
$$\boldsymbol{\alpha}^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

- ▶ Les valeurs estimées

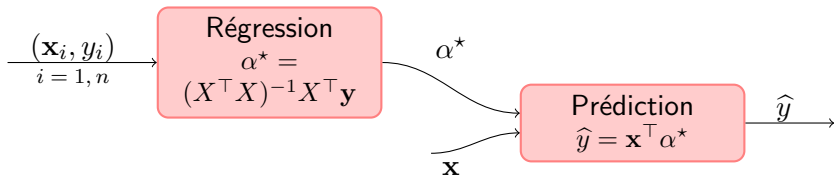
$$\mathbf{z} = X\boldsymbol{\alpha}^* = X(X^\top X)^{-1} X^\top \mathbf{y} = H\mathbf{y}$$

- ▶ Les résidus

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{z} = \mathbf{y} - H\mathbf{y} = (I - H)\mathbf{y}$$

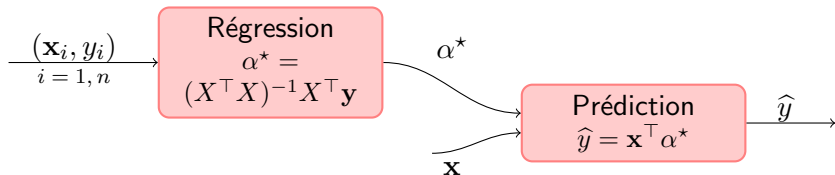


Diagnostic de la régression : les questions



Le diagnostic du modèle : $\hat{y} \pm \delta_y$

Diagnostic de la régression : les questions



Le diagnostic du modèle : $\hat{y} \pm \delta_y$

le modèle le modèle que l'on a posé est-il adapté ?

- ▶ part d'aléa : observation = information + bruit
- ▶ vérifier les hypothèses du modèle : linéaire ?

les observations y a t-il une ou plusieurs observations qui ne conviennent pas ?

- ▶ mauvais x
- ▶ mauvais y
- ▶ mauvais (x, y)

les variables y a t'il une ou plusieurs variables (nuisibles) à éliminer ?

Etude de cas : les données sur le ciment

- ▶ x_1 : Amount of tricalcium aluminate
- ▶ x_2 : Amount of tricalcium silicate
- ▶ x_3 : Amount of tetracalcium alumino ferrite
- ▶ x_4 : Amount of dicalcium silicate
- ▶ y : Heat evolved per gram of cement (in calories)

a = 62.4054 1.5511 0.5102 0.1019 -0.1441
R2 = 0.9824

x1	x2	x3	x4	y	e	h	r	c
7.00	26.00	6.00	60.00	78.50	0.00	0.55	0.00	0.00
1.00	29.00	15.00	52.00	74.30	1.51	0.33	0.71	0.05
11.00	56.00	8.00	20.00	104.30	-1.67	0.58	-0.98	0.26
11.00	31.00	8.00	47.00	87.60	-1.73	0.30	-0.79	0.05
7.00	52.00	6.00	33.00	95.90	0.25	0.36	0.12	0.00
11.00	55.00	9.00	22.00	109.20	3.93	0.12	1.60	0.07
3.00	71.00	17.00	6.00	102.70	-1.45	0.37	-0.70	0.06
1.00	31.00	22.00	44.00	72.50	-3.17	0.41	-1.58	0.34
2.00	54.00	18.00	22.00	93.10	1.38	0.29	0.63	0.03
21.00	47.00	4.00	26.00	115.90	0.28	0.70	0.20	0.02
1.00	40.00	23.00	34.00	83.80	1.99	0.43	1.00	0.15
11.00	66.00	9.00	12.00	113.30	0.97	0.26	0.43	0.01
10.00	68.00	8.00	12.00	109.40	-2.29	0.30	-1.05	0.10

Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

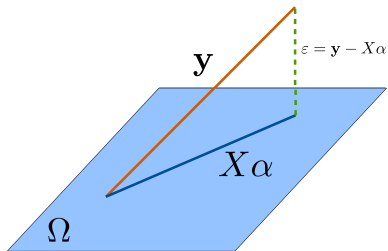
La divergence d'un individu

Les variables sont elles toutes pertinentes

Décomposition de la variance

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCE} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCM}$$

$$\|\mathbf{y} - \bar{y}\mathbf{e}\|^2 = \underbrace{\|\mathbf{y} - \mathbf{z}\|^2}_{\|\boldsymbol{\varepsilon}\|^2} + \|\mathbf{z} - \bar{y}\mathbf{e}\|^2$$



Cette décomposition est valable lorsque le modèle contient une constante.

Relation entre les variables : décomposition de la variance

Posons $z_i = \mathbf{x}_i \alpha^*$, on obtient la décomposition :

$$\begin{aligned}SCT &= \sum_{i=1}^n (y_i - \bar{y})^2 \\&= \sum_{i=1}^n (y_i - z_i + z_i - \bar{y})^2 \\&= \sum_{i=1}^n (y_i - z_i)^2 + \sum_{i=1}^n (z_i - \bar{y})^2 + 2 \underbrace{\sum_{i=1}^n (y_i - z_i)(z_i - \bar{y})}_{=0} \\&= \underbrace{\sum_{i=1}^n (y_i - z_i)^2}_{SCE} + \underbrace{\sum_{i=1}^n (z_i - \bar{y})^2}_{SCM}\end{aligned}$$

Décomposition de la variance :

$$SCT = SCM + SCE$$

observations = modèle + bruit

Degrès de liberté

ddl ?

- ▶ Degré de liberté (ddl) : dimension du sous espace dans lequel se trouve le vecteur
- ▶ Nb d'éléments "libres"
- ▶ Nb de valeurs suffisantes pour déterminer entièrement le vecteur

Source de la variation	nom	ddl
Modèle	SCM	p
Résidus	SCE	$n - 1 - p$
Totale	SCT	$n - 1$

- ▶ p désigne le nombre de variables explicatives
- ▶ Est ce que ce $n - p - 1$ vous dit quelque chose ?

le coefficient de détermination R^2

Définition : le coefficient de détermination R^2

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = \frac{SCM}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

- ▶ R^2 varie entre 0 et 1
- ▶ R^2 mesure l'ajustement sur les données observées
- ▶ Il ne garantit pas que ne garantit pas la validité du modèle, la causalité ou une capacité de prédiction

En python

```
my = np.mean(y)
SCM = np.sum((yp - my)**2)
SCT = np.sum((y - my)**2)
SCE = np.sum((yp - y)**2)
print(SCT, SCE, SCM, SCE+SCM)
```

R^2 pour la régression simple

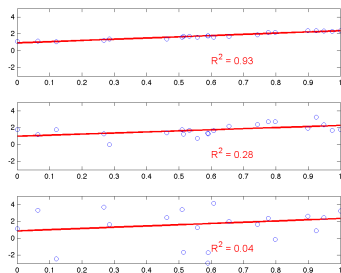
Régression Simple

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (z_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n \mathbf{a}^2 (x_i - \bar{x})^2}{s_y^2} \\ &= \frac{\sum_{i=1}^n \frac{\text{COV}(\mathbf{x}, \mathbf{y})^2}{s_x^4} (x_i - \bar{x})^2}{s_y^2} \\ &= \frac{\text{cov}(\mathbf{x}, \mathbf{y})^2}{s_x^2 s_y^2} = r_{xy}^2 \end{aligned}$$

Trois façons de calculer R^2

```
R_2_cos = (((yp - my)@(y - my))/  
            (np.linalg.norm(y-my) * np.linalg.norm(yp - my)))**2  
R_2_corr = (np.corrcoef(y,yp) [0,1])**2  
R_2_SC = SCM/SCT  
print(R_2_cos,R_2_corr,R_2_SC)
```

le coefficient de détermination R^2 : exemples



- ▶ $\alpha^* = 1,2$ $b^* = 0,95$
- ▶ R^2 différents

Interprétation de R^2

- ▶ Dépend de n (le nombre d'observations) et p (le nombre de variables)
- ▶ R^2 proche de 1 signifie que la modèle est bon
- ▶ R^2 proche de 0 c'est que le modèle est inadapté

Définition : Résidus

Dans le cas de la régression simple on a :

$$\varepsilon_i = y_i - (\alpha^* x_i + b^*), \forall i = 1, \dots, n$$

Dans le cas général, le vecteur des résidus est :

$$\varepsilon = \mathbf{y} - X\alpha^*$$

Qu'est ce qu'un bon résidu ?

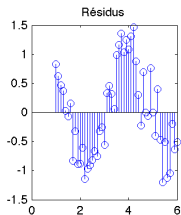
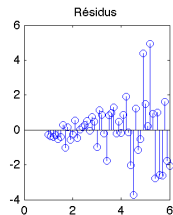
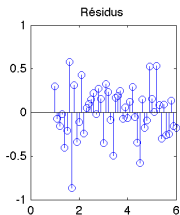
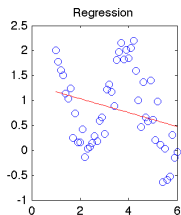
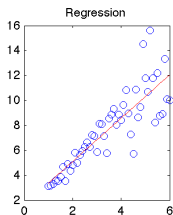
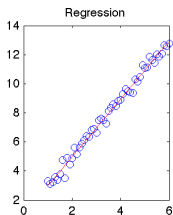
- ▶ Les résidus sont non structurés
- ▶ Ils ne sont pas corrélés avec les variables explicatives.
- ▶ leur distribution est normale
- ▶ Leur variance est constante conditionnellement à X
- ▶ il n'y a pas de point aberrant

Exemples d'Analyse de Résidus

Les différentes figures à examiner sont :

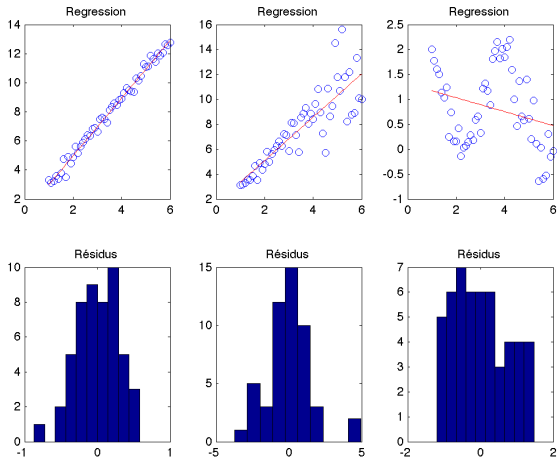
► ε vs x , ε vs y

► Histogramme de ε



Exemples d'Analyse de Résidus

Par hypothèse, les erreurs suivent une loi normale $\mathcal{N}(0, \sigma^2)$. Les résidus doivent être compatibles avec cette hypothèse.



Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

La divergence d'un individu

Les variables sont elles toutes pertinentes

La Contribution des Individus

Contribution

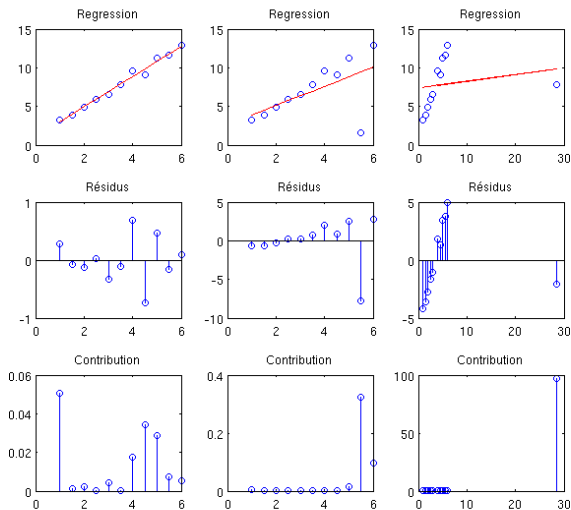
L'influence (la contribution) d'un point se mesure à travers 2 facteurs :

- ▶ Levier $\simeq \mathbf{x}_i - \bar{\mathbf{x}}$
- ▶ Divergence : $y_i - z_i^{\{-i\}}$,
avec $z_i^{\{-i\}}$ la valeur obtenue **SANS** l'observation (\mathbf{x}_i, y_i)

contribution : *combinaison entre levier et divergence*

Différents types de mesures aberrantes

Modèle correct, problème sur une valeur de y , problème sur une valeur de x .



Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

La divergence d'un individu

Les variables sont elles toutes pertinentes

L'influence des x et des y

y valeurs observées / z valeurs estimées

$$\mathbf{y} = \underbrace{X\alpha^*}_{\mathbf{z}} + \varepsilon, \text{ avec } \alpha^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

$$\begin{aligned} \underbrace{\mathbf{z}}_{\text{estimation}} &= X \alpha^* \\ &= \underbrace{X (X^\top X)^{-1} X^\top}_{\text{ce terme ne dépend que de } X} \mathbf{y} = H \mathbf{y} \end{aligned}$$

Influence des $x_i \Leftrightarrow$ Lignes de H

$$z_i = H(i, :) \mathbf{y} = \sum_{j=1}^n H_{ij} y_j, \quad \forall i = 1, \dots, n$$

avec H la matrice dite d'influence $n \times n$:

$$H = X (X^\top X)^{-1} X^\top$$

Matrice d'Influence et Effet Levier

Définition : Effet levier de l'observation i

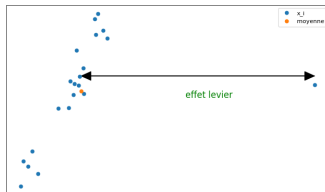
En considérant $z_i = H(i, :)\mathbf{y}$, L'influence de l'observation i , appelée effet levier, est mesurée par :

$$\text{L'Effet levier de l'observation } i = \|H(i, :)\|^2 = H_{ii}$$

NB : $HH = H$, $H^T = H$

Dans le cas de la régression simple ($p=1$)

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\hat{\sigma}_x^2}$$



L'effet levier

Propriétés

$$\sum_{i=1}^n H_{ii} = p + 1 \qquad \frac{1}{n} \leq H_{ii} \leq 1$$

Règle d'usage

Un point x_i à un effet levier important si

$$H_{ii} > 2(p + 1)/n$$

si on faisait l'hypothèse que toutes les variables ont la même influence, tous les $H_{ii} = (p + 1)/n$. En d'autres termes, en moyenne $H_{ii} = (p + 1)/n$. Un point admettent un levier dépassant deux fois sa moyenne (trois fois pour les petits échantillons) est suspect.

Une observation avec un H_{ii} proche de 1 est une observation avec un levier extrêmement important.

Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

La divergence d'un individu

Les variables sont elles toutes pertinentes

La Divergence d'un Individu

Importance d'une observation ?

- ▶ Mesurer comment le modèle évolue lorsque l'on retire chacune des observations.
- ▶ Attention : les points aberrants ont tendance à tirer la droite de régression vers eux !
- ▶ Soit $X_{\{-i\}}$ la matrice des observations sans \mathbf{x}_i
- ▶ $z_i^{\{-i\}}$ la prédiction de \mathbf{x}_i à partir de $X_{\{-i\}}$

Évaluation de l'erreur sans x_i

Définition : Les résidus de validation croisée

$$\varepsilon_i^{\{-i\}} = y_i - \mathbf{z}_i^{\{-i\}}$$

Remarques

- ▶ $X^\top X = X_{\{-i\}}^\top X_{\{-i\}} + (\mathbf{x}_i \mathbf{x}_i^\top)$
- ▶ $\mathbf{z}_i^{\{-i\}} = X_{\{-i\}} \alpha_{\{-i\}}^*$
- ▶ $\alpha_{\{-i\}}^* = \left(X_{\{-i\}}^\top X_{\{-i\}} \right)^{-1} X_{\{-i\}}^\top \mathbf{y}^{\{-i\}}$

Calculer n modèles ?

Fonction $\varepsilon^{\{-i\}} \leftarrow \text{Validation_crois e}(X, \mathbf{y})$

Pour $i = 1, n$ **faire**

1. Construction des donn es : $X_{\{-i\}}$ et $\mathbf{y}^{\{-i\}}$
2. Estimation du mod le
$$\alpha_{\{-i\}}^* = \left(X_{\{-i\}}^\top X_{\{-i\}} \right)^{-1} X_{\{-i\}}^\top \mathbf{y}^{\{-i\}}$$
3. Estimation de l'erreur $\varepsilon_i^{\{-i\}} = y_i - \mathbf{x}_i^\top \alpha_{\{-i\}}^*$

Fin Pour

Faut-il calculer n modèles ? NON !

Théorème : Les Résidus Normalisés

$$\varepsilon_i^{\{-i\}} = \frac{\varepsilon_i}{1 - H_{ii}}$$

Lemme : L'estimateur de validation croisée

$$\alpha_{\{-i\}}^* = \alpha^* - \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1 - H_{ii}}$$

NB : Les ε_i constituent une estimation sans biais de la qualité de prévision d'un modèle car une même observation n'est pas utilisée à la fois pour estimer le modèle et l'erreur de prévision.

Éléments de Preuve

$$1. \alpha_{\{-i\}}^* = \alpha^* - \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1 - H_{ii}} :$$

$$\left(X_{\{-i\}}^\top X_{\{-i\}} \right)^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1} \mathbf{x}_i \mathbf{x}_i^\top (X^\top X)^{-1}}{1 - H_{ii}}$$

avec $H_{ii} = \mathbf{x}_i^\top (X^\top X)^{-1} \mathbf{x}_i$.

La démonstration de ce résultat utilise la formule de Sherman-Morrison qui stipule que

$$(A + uu^\top)^{-1} = A^{-1} + \frac{A^{-1}uu^\top A^{-1}}{1 + u^\top A^{-1}u}$$

on l'utilise avec $A = X^\top X$.

$$2. \varepsilon_i^{\{-i\}} = \frac{\varepsilon_i}{1 - H_{ii}} :$$

$$\begin{aligned} \varepsilon_i^{\{-i\}} &= y_i - \widehat{y}_i^{\{-i\}} \\ &= y_i - \mathbf{x}_i^\top \alpha_{\{-i\}}^* \\ &= y_i - \mathbf{x}_i^\top \left(\alpha^* - \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1 - H_{ii}} \right) \\ &= y_i - \mathbf{x}_i^\top \alpha^* + \frac{\mathbf{x}_i^\top (X^\top X)^{-1} \mathbf{x}_i \varepsilon_i}{1 - H_{ii}} \\ &= \varepsilon_i + \frac{H_{ii} \varepsilon_i}{1 - H_{ii}} = \frac{\varepsilon_i}{1 - H_{ii}} \end{aligned}$$

Pour plus de détails voir :

- ▶ R. Christiansen "Plane answers to complex questions : the theory of linear models" Springer, 2002, p 360
- ▶ p. 291 de Applied Linear Regression, Weisberg

Calcul des Contributions (Distance de Cook)

Distance entre $\alpha_{\{-i\}}^*$ et α^*

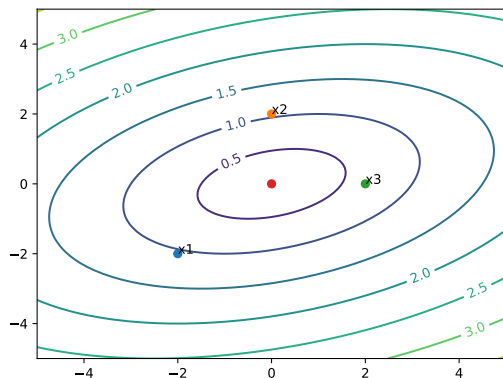
- ▶ Grande différence \Leftrightarrow Forte influence
- ▶ Mauvaise idée : $\|\alpha_{\{-i\}}^* - \alpha^*\|^2$
- ▶ Bonne idée : normaliser \Rightarrow Distance de Mahalanobis

Définition : Distance de Mahalanobis)

Soient \mathbf{x}_1 et \mathbf{x}_2 deux réalisations d'une variable aléatoire gaussienne multidimensionnelle, ayant pour espérance le vecteur μ et pour matrice de variance/covariance Σ . On appelle distance de Mahalanobis entre \mathbf{x}_1 et \mathbf{x}_2 la quantité :

$$d_M(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^\top \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distance de Mahalanobis



- ▶ $d_E(x_1, [0, 0]) \simeq 2,83$; $d_M(\mathbf{x}_1, [0, 0]) \simeq 1,05$
- ▶ $d_E(x_2, [0, 0]) = 2$; $d_M(\mathbf{x}_2, [0, 0]) \simeq 1,05$
- ▶ $d_E(x_3, [0, 0]) = 2$; $d_M(\mathbf{x}_3, [0, 0]) \simeq 0,67$

Analyse de α^*

$$\begin{aligned}\alpha^* &= (X^\top X)^{-1} X^\top \mathbf{y} \\ &= (X^\top X)^{-1} X^\top (X\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) \\ &= (X^\top X)^{-1} X^\top X\boldsymbol{\alpha} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon} \\ &= \boldsymbol{\alpha} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\end{aligned}$$

$$\mathbb{E}(\alpha^*) = \boldsymbol{\alpha}$$

$$\text{car } \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$\begin{aligned}V(\alpha^*) &= V((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\ &= (X^\top X)^{-1} X^\top V(\boldsymbol{\varepsilon}) X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1}\end{aligned}$$

$$\begin{aligned}\text{car } V(\boldsymbol{\varepsilon}) &= \sigma^2 I \\ &= \sigma^2 (X^\top X)^{-1}\end{aligned}$$

$$\alpha^* \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma^2 (X^\top X)^{-1})$$

Contributions (Distance de Cook)

Définition : Contributions (Distance de Cook)

On appelle contribution de l'individu i à la régression, la quantité c_i :

$$c_i = \frac{d_M^2(\alpha_{\{-i\}}^*, \alpha^*)}{p+1} = \frac{(\alpha_{\{-i\}}^* - \alpha^*)^\top X^\top X (\alpha_{\{-i\}}^* - \alpha^*)}{(p+1) s^2}$$

où $\alpha_{\{-i\}}^*$ est le vecteur des coefficients obtenu sans l'exemple (\mathbf{x}_i, y_i) ,
 $p+1$ la dimension du vecteur α^* et $s^2 = \frac{1}{n-p-1} \sum_{i=1}^n \varepsilon_i^2$

Contributions (Distance de Cook)

On peut montrer que les contributions peuvent se réécrire :

$$c_i = \frac{\|\mathbf{z}^{\{-i\}} - \mathbf{z}\|^2}{(p+1) s^2}$$

Pour le calcul pratique on montre que :

Théorème : Calcul pratique de la distance de Cook

$$c_i = \frac{\varepsilon_i^2}{(p+1)s^2} \left[\frac{H_{ii}}{(1-H_{ii})^2} \right]$$

- ▶ Certains auteurs préconisent de se méfier d'une contribution supérieure à un.

Démonstration

Puisque $\alpha_{\{-i\}}^* = \alpha^* - \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1-H_{ii}}$ on a

$$\begin{aligned}(\alpha_{\{-i\}}^* - \alpha^*)^\top X^\top X (\alpha_{\{-i\}}^* - \alpha^*) &= \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1-H_{ii}} X^\top X \frac{(X^\top X)^{-1} \mathbf{x}_i^\top \varepsilon_i}{1-H_{ii}} \\ &= \frac{\varepsilon_i^2}{(1-H_{ii})^2} \mathbf{x}_i (X^\top X)^{-1} \mathbf{x}_i^\top \\ &= \frac{\varepsilon_i^2}{(1-H_{ii})^2} H_{ii}\end{aligned}$$

Tableau de Résultats de la Régression

$$\alpha^* = (X^T X)^{-1} X y$$

$$\varepsilon_i = y_i - \mathbf{x}_i \alpha^* \quad s^2 = \frac{1}{n-p} \sum_{i=1}^n \varepsilon_i^2 \quad H = X(X^T X)^{-1} X^T$$

- ▶ Résidus de validation croisée : $\varepsilon_i^{\{-i\}} = \frac{\varepsilon_i}{1-H_{ii}}$
- ▶ Résidus standardisés : $r_i = \frac{\varepsilon_i}{s\sqrt{1-H_{ii}}}$
- ▶ Résidus studentisés : $t_i = \frac{\varepsilon_i}{s^{\{-i\}}\sqrt{1-H_{ii}}}$

$$c_i = \frac{H_{ii}}{p(1-H_{ii})^2} \frac{\varepsilon_i^2}{s^2}$$

variable explicatives	variables à expliquer	résidus erreurs	résidus de VC	contributions
X	y_i	ε	$\varepsilon^{\{-i\}}$	c

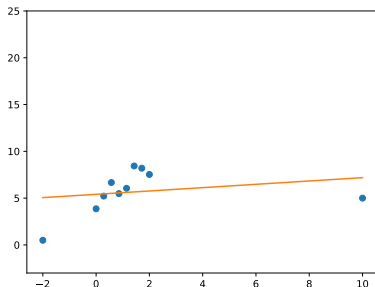
Exemple

	x	y	z	e	e_n	h	c
0	-2	0.893605	4.35849	-3.46488	-4.58485	0.244275	0.798133
1	0	4.19674	4.73943	-0.542688	-0.622705	0.128499	0.00774482
2	0.285714	4.30322	4.79385	-0.490628	-0.557044	0.119229	0.00575055
3	0.571429	5.10903	4.84827	0.260765	0.29358	0.111778	0.00149746
4	0.857143	4.85229	4.90268	-0.0503969	-0.0563814	0.106143	5.24459e-05
5	1.14286	5.96273	4.9571	1.00562	1.12026	0.102326	0.0199604
6	1.42857	7.2051	5.01152	2.19358	2.4382	0.100327	0.0927051
7	1.71429	5.77542	5.06594	0.709478	0.788436	0.100145	0.00967635
8	2	7.14363	5.12036	2.02326	2.25253	0.101781	0.0802704
9	10	5	6.64411	-1.64411	-14.3586	0.885496	28.3765

► x_4, y_4 : Résidu et résidu similaires, Faible contribution

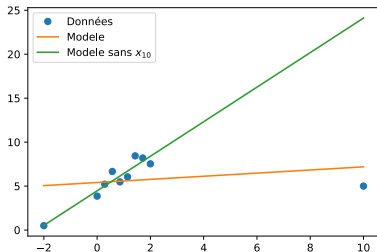
► x_0, y_0 : Résidu élevé, Lever moyen, Contribution moyenne

► x_9, y_9 : Résidu moyen, Fort levier, Forte contribution



Exemple

	x	y	z	e	e_n	h	c
0	-2	0.893605	0.960652	-0.0670465	-0.251425	0.733333	0.0762761
1	0	4.19674	4.02698	0.169761	0.199719	0.15	0.00984468
2	0.285714	4.30322	4.46502	-0.161805	-0.184669	0.12381	0.00694726
3	0.571429	5.10903	4.90307	0.205961	0.231913	0.111905	0.00990308
4	0.857143	4.85229	5.34112	-0.488827	-0.551902	0.114286	0.052728
5	1.14286	5.96273	5.77916	0.183567	0.211227	0.130952	0.0096136
6	1.42857	7.2051	6.21721	0.987896	1.17874	0.161905	0.370141
7	1.71429	5.77542	6.65525	-0.879833	-1.1097	0.207143	0.419713
8	2	7.14363	7.0933	0.0503268	0.0686274	0.266667	0.0020665



Résumons nous

Étant donné des observations (\mathbf{x}_i, y_i) , $\mathbf{x}_i \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$,

1. Construire la matrice des données $X = [\mathbf{e} \ \mathbf{x}]$
2. Calculer les coefficients : $\alpha^* = (X^\top X)^{-1} X \mathbf{y}$
3. Calculer les valeurs prédites : $\mathbf{z} = X \alpha^*$
4. Calculer les résidus : $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{z}$
 - ↪ Si les résidus présentent une structure :
 - ↪ Transformer les données et reprendre à 1.
5. Calculer le R^2
 - ↪ Si R^2 est trop petit la régression n'a pas de sens
6. Calculer la variance estimée : $s^2 = \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{n-1-p}$
7. Calculer les contributions : $\mathbf{c} = \frac{\mathbf{h}}{(p+1)(1-\mathbf{h})^2} \frac{\boldsymbol{\varepsilon}^2}{s^2}$, $\mathbf{h} = \text{diag}(H)$
 - ↪ Si la contribution d'un point est supérieure à $\frac{4}{n}$:
 - ↪ Examiner le point correspondant

Résumons nous : le (pseudo) code qui va bien

- $X = [\mathbf{e} \ \mathbf{x}]$
- $\alpha^* = (X^\top X)^{-1} X \mathbf{y}$ `a = np.linalg.solve((X.T@X), (X.T@y))`
- $\mathbf{z} = X \alpha^*$ `z = X@a`
- $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{z}$ `e = y-z`

- $R^2 = \frac{(z - \bar{y})^\top (z - \bar{y})}{(y - \bar{y})^\top (y - \bar{y})}$ `R2 = 1 - e.T@e/np.sum((y-np.mean(y))**2)`
- $H = X(X^\top X)^{-1} X^\top$ `H = X@(np.linalg.solve((X.T@X), (X.T)))`
 `h = np.diag(H)`
- $\mathbf{h} = \text{diag}(H)$

- $s^2 = \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{n-p}$ `s2 = e.T@e/(n-p-1)`
- $\mathbf{c} = \frac{\mathbf{h}}{p(1-\mathbf{h})^2} \frac{\boldsymbol{\varepsilon}^2}{s^2}$ `c = (h/(1-h)**2)*(e**2) / ((p+1)*s2)`

Diagnostic de la régression

Diagnostic Général

- ▶ R^2 : qualité de la régression
- ▶ Analyse des résidus
 - ▶ Structure
 - ▶ Gaussien
 - ▶ Indépendance des y et x

Analyse des individus

- ▶ Calcul de l'influence des \mathbf{x}_i : $H_{i,i}$
- ▶ Résidus normalisés : $\varepsilon_i^{\{-i\}} = \frac{\varepsilon_i}{1-H_{ii}}$
- ▶ Contribution : $c_i = \frac{H_{ii}}{(p+1)(1-H_{ii})^2} \frac{\varepsilon_i^2}{s^2}$

Diagnostic de la régression

Diagnostic Général

- ▶ R^2 : qualité de la régression
- ▶ Analyse des résidus
 - ▶ Structure
 - ▶ Gaussien
 - ▶ Indépendance des y et x

Analyse des individus

- ▶ Calcul de l'influence des x_i : $H_{i,i}$
- ▶ Résidus normalisés : $\varepsilon_i^{\{-i\}} = \frac{\varepsilon_i}{1-H_{ii}}$
- ▶ Contribution : $c_i = \frac{H_{ii}}{(p+1)(1-H_{ii})^2} \frac{\varepsilon_i^2}{s^2}$

Reste à faire : les **variables**

Plan

Diagnostic de la régression

Les objectifs de l'analyse du modèle

Qualité du modèle

Y a-t'il une relation entre les variables

La relation est elle linéaire : l'examen des résidus

Y a-t'il des individus hors épure

La contribution d'un individu

La matrice d'influence

La divergence d'un individu

Les variables sont elles toutes pertinentes

La pertinence des variables

$$y = a_0 + \sum_{j=1}^p a_j \mathbf{x}(j) + \varepsilon$$

La problématique

- ▶ Supposons que $a_j = 0$
- ▶ $\Rightarrow \mathbf{x}(j)$ n'intervient pas dans la régression
- ▶ Que faire si X contient de “mauvais” prédicteurs ?

La sélection de variables

- ▶ Trouver le “meilleur” sous-ensemble de variables p_0
- ▶ p_0 petit :
 - ▶ Meilleur conditionnement de $X^\top X$
 - ▶ Calcul plus rapides
 - ▶ Modèle plus simple (rasoir d'Ockham)

Stratégies de sélection

Stratégie globale

- ▶ Tester différents sous ensembles de variables
- ▶ Garder le “meilleur” sous-ensemble donné un critère
 - ▶ Lequel ?

Stratégies de parcours des sous ensembles de variables :

- ▶ Approche exhaustive : $2^p - 1$ modèles potentiels
- ▶ Forward/Backward
- ▶ Stepwise
- ▶ ...

Etude de cas : les données sur le ciment

- ▶ x_1 : Amount of tricalcium aluminate
- ▶ x_2 : Amount of tricalcium silicate
- ▶ x_3 : Amount of tetracalcium alumino ferrite
- ▶ x_4 : Amount of dicalcium silicate
- ▶ y : Heat evolved per gram of cement (in calories)

a = 62.4054 1.5511 0.5102 0.1019 -0.1441
R2 = 0.9824

x1	x2	x3	x4	y	e	h	r	c
7.00	26.00	6.00	60.00	78.50	0.00	0.55	0.00	0.00
1.00	29.00	15.00	52.00	74.30	1.51	0.33	0.71	0.05
11.00	56.00	8.00	20.00	104.30	-1.67	0.58	-0.98	0.26
11.00	31.00	8.00	47.00	87.60	-1.73	0.30	-0.79	0.05
7.00	52.00	6.00	33.00	95.90	0.25	0.36	0.12	0.00
11.00	55.00	9.00	22.00	109.20	3.93	0.12	1.60	0.07
3.00	71.00	17.00	6.00	102.70	-1.45	0.37	-0.70	0.06
1.00	31.00	22.00	44.00	72.50	-3.17	0.41	-1.58	0.34
2.00	54.00	18.00	22.00	93.10	1.38	0.29	0.63	0.03
21.00	47.00	4.00	26.00	115.90	0.28	0.70	0.20	0.02
1.00	40.00	23.00	34.00	83.80	1.99	0.43	1.00	0.15
11.00	66.00	9.00	12.00	113.30	0.97	0.26	0.43	0.01
10.00	68.00	8.00	12.00	109.40	-2.29	0.30	-1.05	0.10

Sélection systématique des variables

1	x1	x2	x3	x4	R ²	
1	x1				0.534	$y = a_0 + a_1x_1$
1		x2			0.666	...
1			x3		0.286	
1				x4	0.675	$y = a_0 + a_4x_4$
1	x1	x2			0.979	$y = a_0 + a_1x_1 + a_2x_2$
1	x1		x3		0.548	
1	x1			x4	0.972	...
1		x2	x3		0.847	$y = a_0 + a_3x_3 + a_4x_4$
1		x2		x4	0.680	...
1			x3	x4	0.935	...
1	x1	x2	x3		0.982	$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$
1	x1	x2		x4	0.982	
1	x1		x3	x4	0.982	...
1		x2	x3	x4	0.973	$y =$
1	x1	x2	x3	x4	0.982	$a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$

Quelles variables choisir ?

$$\{x_1, x_2\} : R^2 = 0.979$$

$$\{x_1, x_2, x_3, x_4\} : R^2 = 0.982$$

Le R^2 n'est pas suffisant pour choisir...

Mesure de la qualité d'un modèle : le C_p de Mallows

Définition : C_p de Mallows

Soit $X^{(0)} \in \mathbb{R}^{n \times p_0}$ une sous matrice de $X \in \mathbb{R}^{n \times p}$, avec $p_0 < p$ et $z_i^{(0)}$ le vecteur des prédictions associées.

Le C_p de Mallows de la matrice $X^{(0)}$ est défini par :

$$C_p = \frac{1}{s^2} \sum_{i=1}^n (y_i - z_i^{(0)})^2 - n + 2p_0$$

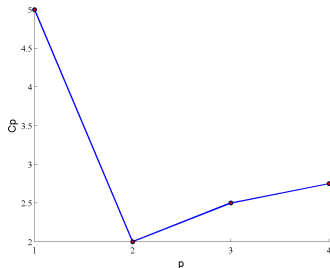
avec $s^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - z_i)^2$ la variance estimée sur le modèle complet .

NB : Ici, p désigne le nombre total de colonnes de X utilisé pour la régression, incluant donc la colonne de 1.

Analyse du C_p de Mallows

$$C_p = \frac{1}{s^2} \sum_{i=1}^n (y_i - z_i^{(0)})^2 - n + 2p_0$$

- ▶ le C_p estime l'erreur quadratique
- ▶ plus le C_p d'un modèle est petit, meilleur est ce modèle
 - ▶ plus il y a de variables, plus $\sum_{i=1}^n (y_i - z_i^{(0)})^2$ est petit
 - ▶ mais p_0 augmente



Sélection systématique des variables

1	x1	x2	x3	x4	R ²	erreurs	Cp
1	x1				0.534	1699.61	202.55
1		x2			0.666	1202.09	142.49
1			x3		0.286	2616.36	315.15
1				x4	0.675	1194.22	138.73
1	x1	x2			0.979	93.88	2.68
1	x1		x3		0.548	2218.12	198.09
1	x1			x4	0.972	121.22	5.50
1		x2	x3		0.847	701.74	62.44
1		x2		x4	0.680	1461.81	138.23
1			x3	x4	0.935	294.01	22.37
1	x1	x2	x3		0.982	90.00	3.04
1	x1	x2		x4	0.982	85.35	3.02
1		x2	x3	x4	0.973	146.85	7.34
1	x1	x2	x3	x4	0.982	110.35	5.00

Quelles variables choisir ? \min_{Cp}

$$\{x_1, x_2\} : Cp = 2.68$$

Sélection de variables : autres approches

Forward Selection

► Approche additive

1. Ensemble de départ vide
2. À chaque itération, on ajoute la variable donnant une matrice $X^{(0)}$ avec un C_p minimum
3. Quand toutes les variables ont été ajoutées, on prend le min des C_p sur l'ensemble.

Backward Selection

► Approche soustractive

1. Ensemble de départ avec **toutes les variables**
2. À chaque itération, on **supprime** la variable donnant une matrice $X^{(0)}$ avec un C_p minimum
3. Quand toutes les variables ont été **supprimées**, on prend le min des C_p sur l'ensemble.

Conclusion

Diagnostic

- ▶ Modèle :
 - ▶ R^2
 - ▶ Examen global des résidus
- ▶ Observations
 - ▶ Examen des résidus un à un
 - ▶ Examen des contributions (distance de Cook)
- ▶ Variables
 - ▶ calcul du C_p de Mallows

Pour aller plus loin

- ▶ Critères BIC, AIC, PRESS, ...
- ▶ Utiliser des tests statistiques
- ▶ Changer de fonction objectif

$$\sum (y_i - z_i)(z_i - \bar{y}) = 0?$$

$$\begin{aligned}\sum (y_i - z_i)(z_i - \bar{y}) &= \sum \varepsilon_i (z_i - \bar{y}) \\ &= \sum \varepsilon_i \bar{y} - \varepsilon_i z_i \\ &= \bar{y}n \underbrace{\bar{\varepsilon}}_{=0} - \underbrace{\varepsilon^\top \mathbf{z}}_{=0}\end{aligned}$$

$$\bar{y} = \bar{z}$$

- ▶ $y \in \mathbb{R}^n$: vecteur des observations
- ▶ $z = \hat{y} = X\hat{\alpha}$: prédictions du modèle
- ▶ $\hat{\alpha} = (X^\top X)^{-1}X^\top y$
- ▶ $H = X(X^\top X)^{-1}X^\top$: matrice de projection

Montrer que $\bar{z} = \bar{y}$

$$\begin{aligned}\bar{z} &= \frac{1}{n}\mathbf{1}^\top z = \frac{1}{n}\mathbf{1}^\top Hy \\ &= \frac{1}{n}(H^\top \mathbf{1})^\top y \quad (\text{car } H \text{ est symétrique}) \\ &= \frac{1}{n}\mathbf{1}^\top y \quad (\text{si } \mathbf{1} \in \text{Im}(X) \Rightarrow H\mathbf{1} = \mathbf{1}) \\ &= \bar{y}\end{aligned}$$