Diagnostic de la régression

Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

May 17, 2011

Plan

Le modèle de régression (rappels)

- Les données : $(\mathbf{x}_i, y_i)$ (n observations)
- Le Modèle ($p+1$ inconnues)

$$\mathbf{y} = X\alpha + \varepsilon \qquad y = a_0 + a_1 x_1 + \cdots + a_p x_p + \varepsilon$$

- Le principe de projection

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

- Les coefficients de la régression

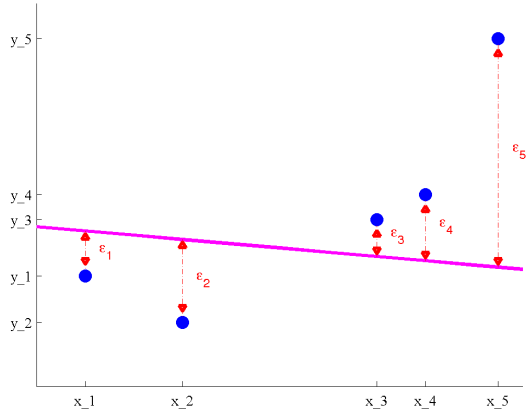
$$\widehat{\alpha} = (X^\top X)^{-1} X^\top \mathbf{y}$$

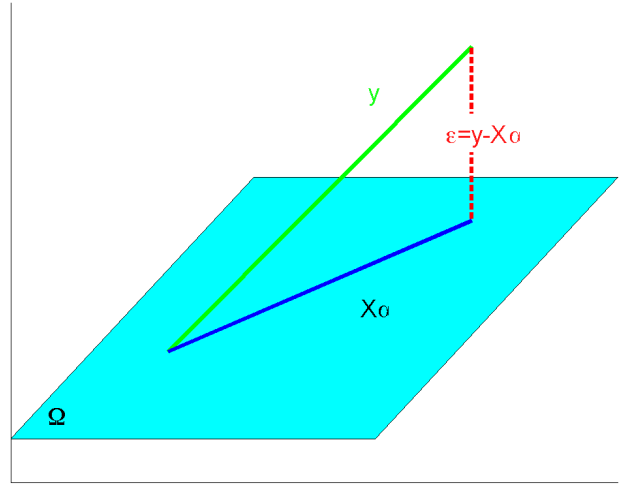
- Les valeurs estimées

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

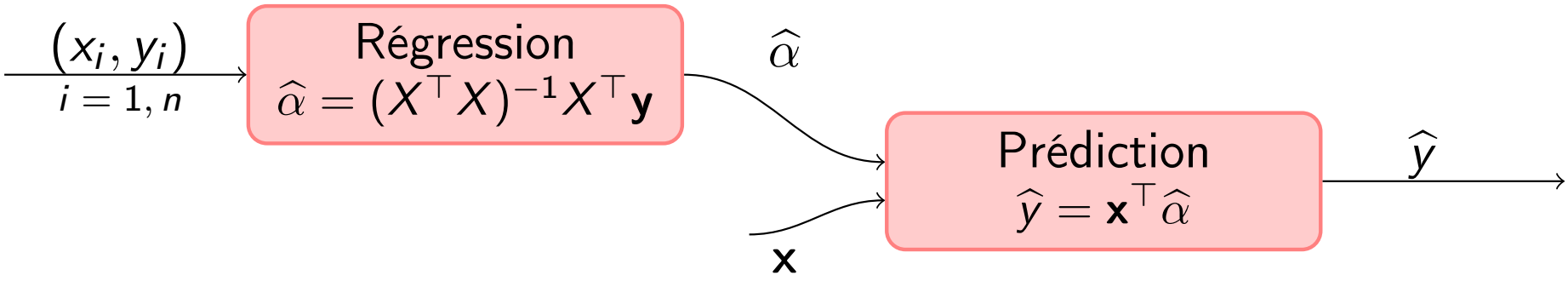
- Les résidus

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



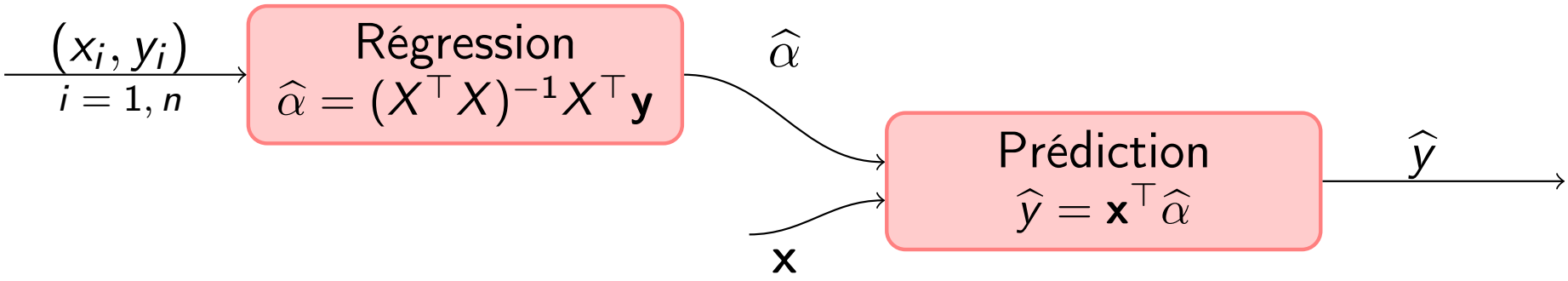


Diagnostic de la régression : les questions



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

Diagnostic de la régression : les questions



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

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

- part d'aléat : 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 ?

Plan

Pourquoi le modèle peut être mauvais

- la part de bruit est trop importante
  - part d'aléat : observation = information + bruit
- Le modèle est mal spécifié
  - soit il n'y a globalement pas de relation entre les variables explicatives et la variable à expliquer,
  - soit cette relation n'est pas linéaire.
- Les hypothèses du modèle ne sont pas vérifiées
  - toutes les observations sont analogues

Relation entre les variables : décomposition de la variance

Si l'on pose $z_i = \mathbf{x}_i\widehat{\alpha} \quad (= \widehat{a}x_i + \widehat{b})$ 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 :

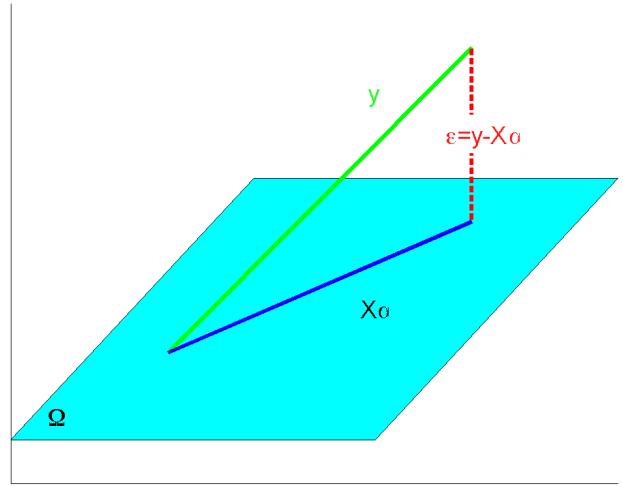
$$SCT = SCM + SCE$$

observations = modèle + bruit

Décomposition de la variance

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

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



Source de la variation	nom	ddl	Moyenne
modèle	SCM	p	$\frac{SCM}{p}$
résidus	SCE	$n-(p+1)$	$\frac{SCE}{n-1-p}$
totale	SCT	$n-1$	$\frac{SCT}{n-1}$

Table: Tableau d'analyse de la variance

degré de liberté(ddl) = dimension du sous espace dans lequel se trouve le vecteur

le coefficient de détermination R^2

Definition (le coefficient de détermination R^2)

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

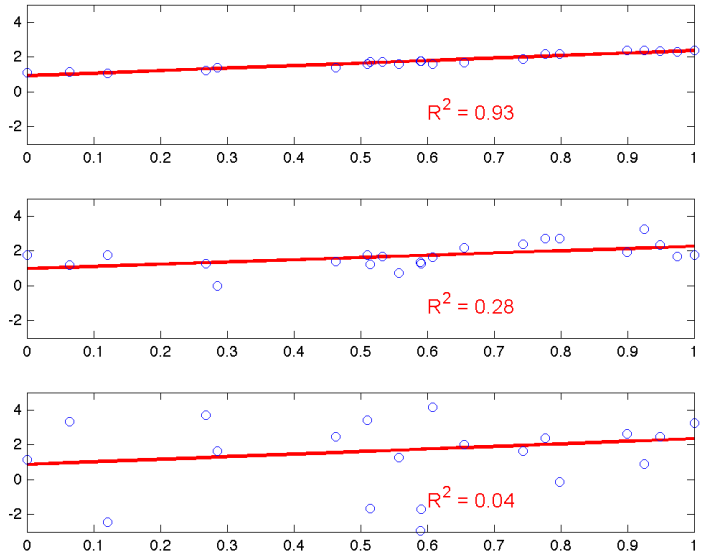
pour la régression simple

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

- R^2 varie entre 0 et 1
- R^2 proche de 1 signifie que la modèle est bon
- R^2 proche de 0 c'est que le modèle est inadapté

le coefficient de détermination R^2 : exemples

régression simple



régression multiple

$\widehat{a} = 1,2 \quad \widehat{b} = 0,95$

L'interprétation de R^2 dépend de n (le nombre d'observations)

- R^2 proche de 1 signifie que la modèle est bon
- R^2 proche de 0 c'est que le modèle est inadapté

La relation est elle linéaire : l'examen des résidus

Definition (les résidus)

Dans le cas de la régression simple on a : $\widehat{\varepsilon}_i = y_i - \left(\widehat{a}x_i + \widehat{b}\right) \quad i = 1, n$
Dans le cas général, le vecteur des résidus est :

$$\widehat{\varepsilon} = \mathbf{y} - X\widehat{\alpha}$$

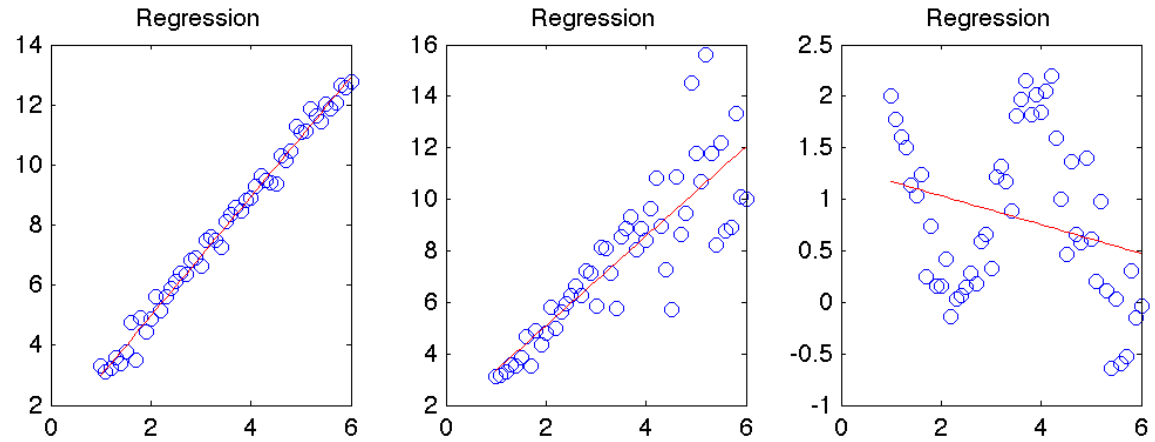
Un bon résidu est un résidu sans structure ou sans structure apparente...

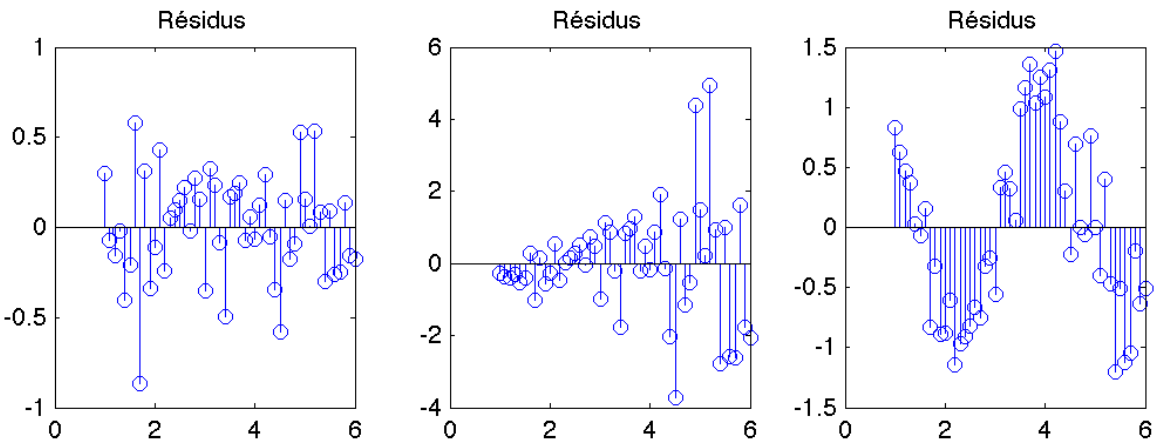
- les résidus non structurés
  - leur variance est constante
  - ils sont indépendant des observations ($\mathbf{x}$ et y)
- leur distribution est normale
- il n'y a pas de point aberrant

Les différentes figures à examiner sont :

- $\widehat{\varepsilon}$ vs $\mathbf{x}$, $\widehat{\varepsilon}$ vs $\mathbf{y}$
- histogramme des $\widehat{\varepsilon}$

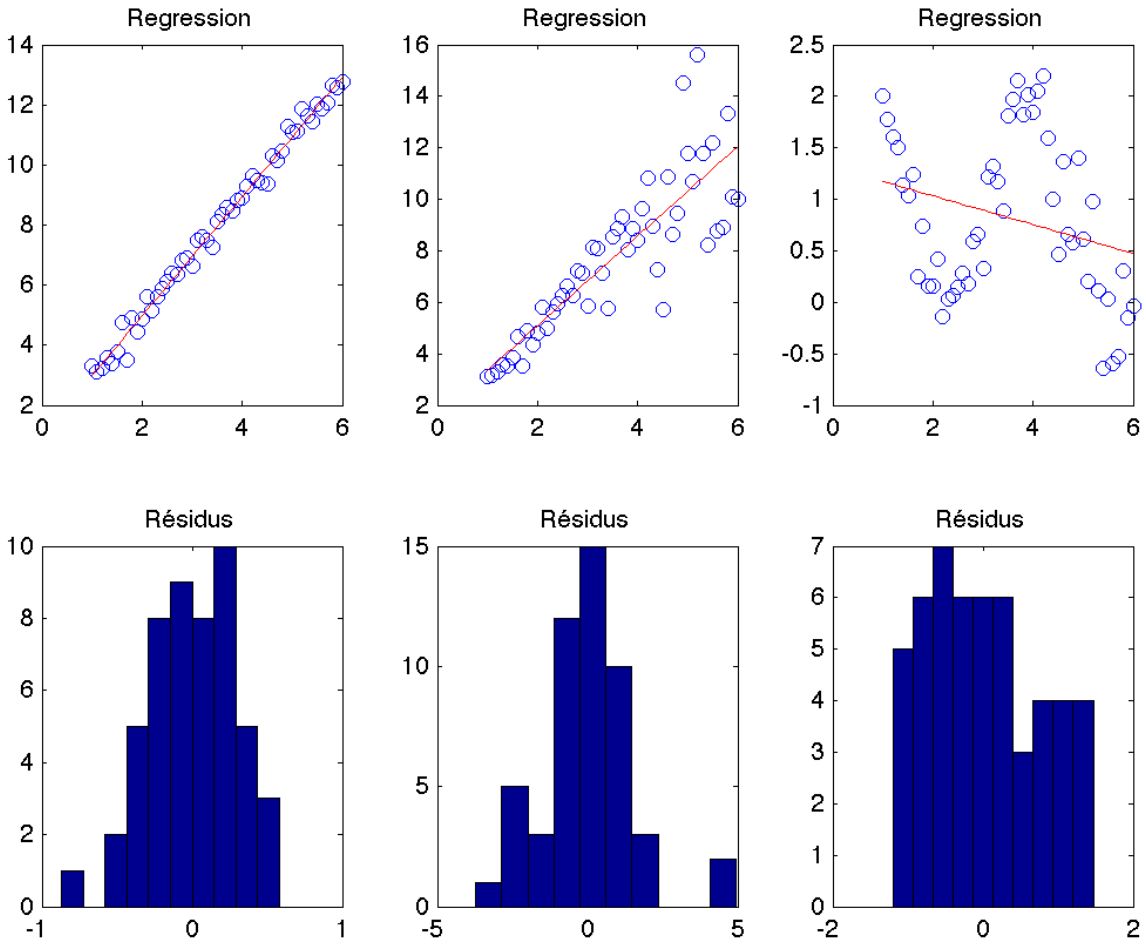
Exemples de pathologies des résidus





Différents types de problèmes dans la régression. A gauche le modèle est correct, au centre la variance est non constante et à droite le modèle n'est pas linéaire

Exemples de pathologies des résidus



quand le modèle est bon, la distribution des résidus est normale (loi gaussiène)

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  = 62.40  1.55  0.51  0.10  -0.14
R2 = 0.98
```

x1	x2	x3	x4	y	e
7.00	26.00	6.00	60.00	78.50	0.00
1.00	29.00	15.00	52.00	74.30	1.51
11.00	56.00	8.00	20.00	104.30	-1.67
11.00	31.00	8.00	47.00	87.60	-1.73
7.00	52.00	6.00	33.00	95.90	0.25
11.00	55.00	9.00	22.00	109.20	3.93
3.00	71.00	17.00	6.00	102.70	-1.45
1.00	31.00	22.00	44.00	72.50	-3.17
2.00	54.00	18.00	22.00	93.10	1.38
21.00	47.00	4.00	26.00	115.90	0.28
1.00	40.00	23.00	34.00	83.80	1.99
11.00	66.00	9.00	12.00	113.30	0.97
10.00	68.00	8.00	12.00	109.40	-2.29

Il faut aussi afficher les résidus : $\widehat{\varepsilon}$ vs $\mathbf{y}$

Exemple de transformation des données

La relation de Faber-Jackson est une loi empirique qui à la forme suivante :

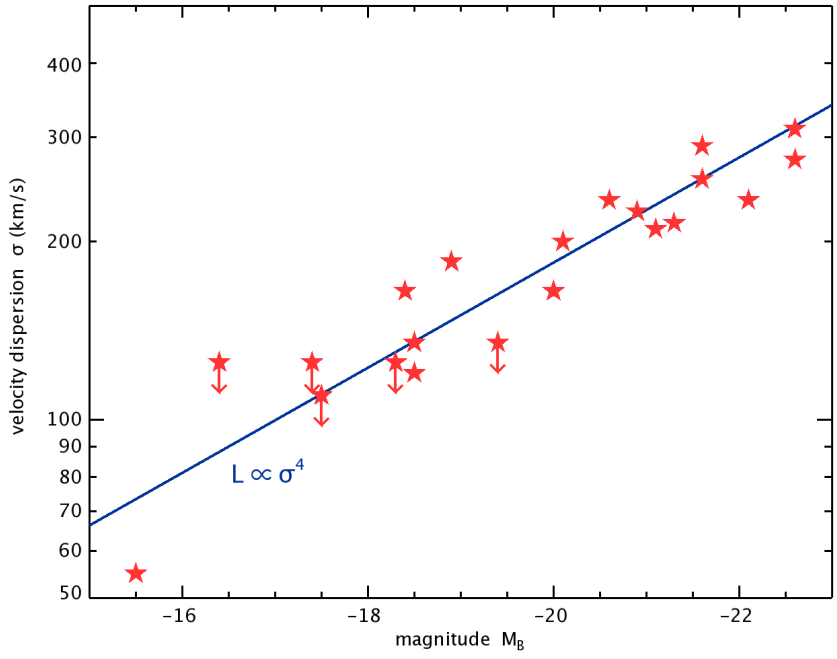
$$L = k\sigma^{\beta}$$

- σ vitesse d'expansion dans les galaxies elliptiques, mesurée à partir de l'observation de la vitesse d'émission de gaz,
- L luminosité apparente,
- β un index proche de 4.

On en déduit le modèle linéaire :

$$\underbrace{\log L}_{y} = \underbrace{\beta}_{a}\underbrace{\log \sigma}_{x} + \underbrace{\log k}_{b}$$

vérifier si oui ou non $\beta = 4$?



Exemples de données qui semblent linéairement liées

`http://www.math.yorku.ca/SCS/Gallery/`

Plan

La contribution des individus

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

levier $x_i - \bar{x}$

divergence $y_i - z_i^{-i}$
avec z_i^{-i} la valeur obtenue SANS l'observation $(\mathbf{x}_i, y_i)$

contribution = levier $\times$ divergence

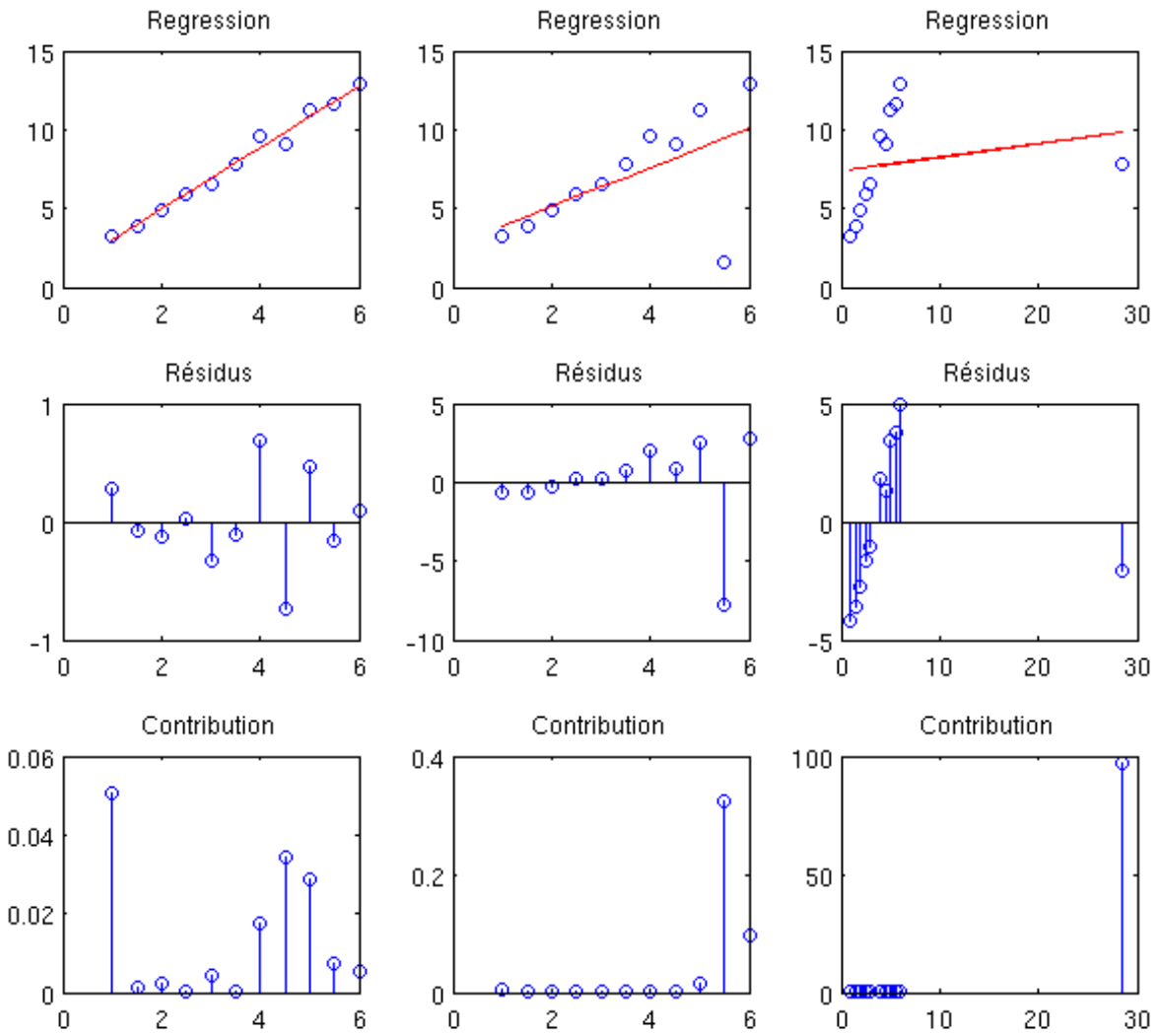


Figure: Différents types de mesures aberrantes. A gauche le modèle est correct, au centre une valeur de y est mauvaise et a droite c'est une mesure sur x qui est aberrante.

Plan

L'influence des x et des y

y valeurs observées / z valeurs estimées

$$\mathbf{y} = \underbrace{X\widehat{\alpha}}_{\mathbf{z}} + \widehat{\varepsilon}$$

avec

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

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

l'influence des x : les lignes de la matrice H

$$z_i = H_{i\bullet}^\top \mathbf{y} = \sum_{j=1}^{n} H_{ij} y_j \qquad i = 1, n$$

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

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

la matrice d'influence et l'effet levier

Definition (influence des X : l'effet levier de l'observation i)

Puisque $\mathbf{z} = H\mathbf{y}$, pour chacune de ses composantes on a $z_i = H_{i\bullet}^{\top}\mathbf{y}$
L'influence de l'observation i est mesurée par la norme du vecteur $H_{i\bullet}$:

$$\text{l'effet levier de l'observation } i = \|H_{i\bullet}\|^2 = \sum_{j=1}^{n} H_{ij}^2$$

H est aussi une matrice de projection ($H^2 = H$) symétrique ($H^{\top} = H$).

Theorem (La diagonale de la matrice H : H_{ii})

L'influence de l'observation i dans le calcul de la régression est donnée par le terme diagonal H_{ii} de la matrice d'influence H car $\|H_{i\bullet}\|^2 = H_{ii}$

Démonstration : Puisque $H^2 = H$ et que H est symétrique ($H^{\top} = H$)

$$H_{ii} = \sum_{j=1}^{n} H_{ij}H_{ji} = \sum_{j=1}^{n} H_{ij}H_{ij} = \sum_{j=1}^{n} H_{ij}^2 = \|H_{i\bullet}\|^2$$

Propriétés: $trace(H) = dim(\Omega) = p + 1$; $0 \leq H_{ii} \leq 1$; $H_{ii} = \mathbf{x}_i(X^{\top}X)^{-1}\mathbf{x}_i^{\top}$

Matrice d'influence, effet levier et distance à la moyenne

Theorem (Dans le cas de la régression simple)

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Démonstration : $z_i = \widehat{a}x_i + \widehat{b} = \widehat{a}x_i + (\bar{y} - \widehat{a}\bar{x}) = \bar{y} + \widehat{a}(x_i - \bar{x})$

$$\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{n} y_j + \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}(x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{j=1}^{n} y_j + \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}(x_i - \bar{x}) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\right) y_j
\end{aligned}$$

De manière générale on peut réécrire les leviers en fonction de $\mathbf{x}_i - \bar{x}$

$$H_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{x})(X^\top X)^{-1}(\mathbf{x}_i - \bar{x})^\top$$

où $\mathbf{x}_i$ est la i ème ligne de la matrice X et $\bar{x}$ le vecteur des moyennes

La matrice d'influence H ne dépend que des X

```
X = [ones(5,1) sort(rand(5,1)); 1 5] =

    1.0000    0.0305
    1.0000    0.4799
    1.0000    0.5000
    1.0000    0.7441
    1.0000    0.9047
    1.0000    5.0000

H = X*inv(X'*X)*X'  =

    0.2576    0.2248    0.2233    0.2055    0.1938   -0.1050
    0.2248    0.2038    0.2029    0.1915    0.1840   -0.0070
    0.2233    0.2029    0.2020    0.1909    0.1836   -0.0027
    0.2055    0.1915    0.1909    0.1833    0.1783    0.0506
    0.1938    0.1840    0.1836    0.1783    0.1748    0.0856
   -0.1050   -0.0070   -0.0027    0.0506    0.0856    0.9786

diag(H)  =       0.2576    0.2038    0.2020    0.1833    0.1748    0.9786

x = X(:,2);
1/(n+1)  +  (x - mean(x)).^2/((x - mean(x))'*(x - mean(x)))
         =       0.2576    0.2038    0.2020    0.1833    0.1748    0.9786

sum(diag(H)) =   2
```

L'effet levier

Propriétés des leviers :

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

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.

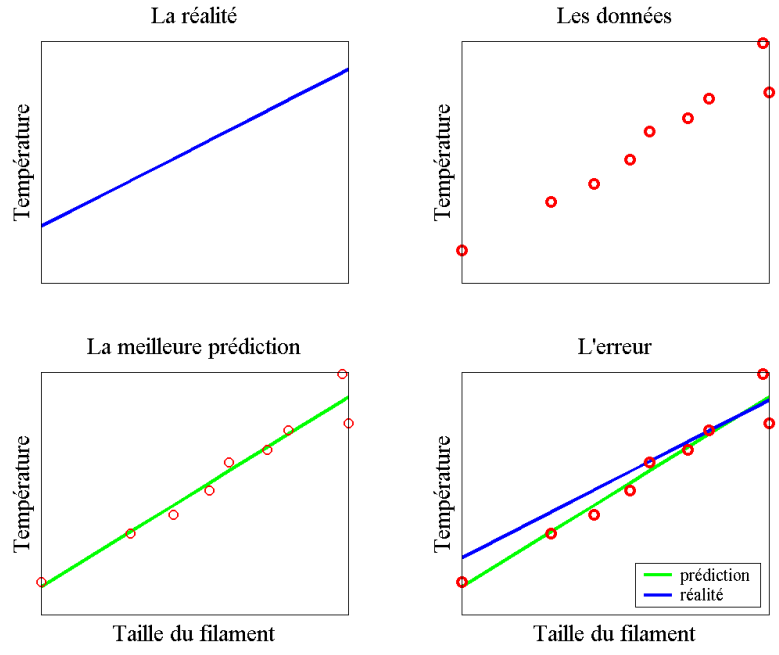
Plan

La matrice d'influence et les résidus

$$
\begin{aligned}
z &= H\mathbf{y};\\
H &= X(X^\top X)^{-1}X^\top
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\varepsilon} &= \mathbf{y} - z\\
&= (I - H)\mathbf{y};
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\varepsilon} &= (I - H)\mathbf{y}\\
&= (I - H)\left(X\mathbf{a} + \varepsilon\right)\\
&= (I - H)X\mathbf{a} + (I - H)\varepsilon \quad = (I - H)\varepsilon
\end{aligned}
$$



car : $(I - H)X = X - HX = X - (X(X^\top X)^{-1}X^\top)X = X - X = 0$

Standatdisation des résidus

$$\begin{aligned}
\widehat{\varepsilon} &= (I-H)\varepsilon \\
\mathbb{E}(\widehat{\varepsilon}) &= \mathbb{E}\big((I-H)\varepsilon\big) = (I-H)\mathbb{E}(\varepsilon) = 0 \qquad \text{si par hypothèse } \mathbb{E}(\varepsilon)=0 \\
V(\widehat{\varepsilon}) &= V\big((I-H)\varepsilon\big) \\
&= (I-H)V(\varepsilon)(I-H)^\top \\
&= \sigma^2(I-H)(I-H)^\top \qquad \text{si par hypothèse } V(\varepsilon)=\sigma^2 I \\
&= \sigma^2(I-H-H+H*H) = \sigma^2(I-H)
\end{aligned}$$

$$V(\widehat{\varepsilon}) = \sigma^2(I-H) \quad \Leftrightarrow \begin{cases} V(\widehat{\varepsilon}_i) & = \sigma^2(1-H_{ii}) \\ cov(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j) & = -\sigma^2 H_{ij} \end{cases}$$

D'où l'idée de normaliser les résidu par rapport à leur écart type.

Definition (résidus standardisés r_i)

$$r_i = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{V(\widehat{\varepsilon}_i)}}} = \frac{\widehat{\varepsilon}_i}{s\sqrt{1-H_{ii}}}$$

avec $s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$

Plan

La divergence d'un individu

- Attention : les points aberrants ont tendance à tirer la droite de régression vers eux !
- mesurer comment le modèle évolue lorsque l'on retire chacun des observations.

Dans ce cas, si l'on note $\mathbf{x}_i$ la i ème observation, $X_{(-i)}$ la matrice des observations sans cette i ème observation et $z_i^{(-i)}$ la prédiction réalisé au point $\mathbf{x}_i$ par un modèle identifié sans la i ème observation,

Definition (Les résidus de validation croisée)

$$\widehat{\varepsilon}_i^{(-i)} = y_i - z_i^{(-i)}$$

Remarque : $X^\top X = X_{(-i)}^\top X_{(-i)} + (\mathbf{x}_i^{(-i)})^\top \mathbf{x}_i^{(-i)}$; $z_i^{(-i)} = X_{(-i)}\, \widehat{\alpha}^{(-i)}$ et

$$\widehat{\alpha}^{(-i)} = \left(X_{(-i)}^\top X_{(-i)}\right)^{-1} X_{(-i)}^\top \mathbf{y}^{(-i)}$$

Calculer n modèles ?

Fonction $\widehat{\varepsilon}^{(-i)} \leftarrow$ `Validation_croisée`$(X, \mathbf{y})$

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

1. construction des données : $X_{(-i)}$ et $\mathbf{y}^{(-i)}$
2. estimation du modèle $\widehat{\alpha}^{(-i)} = \left(X_{(-i)}^\top X_{(-i)}\right)^{-1} X_{(-i)}^\top \mathbf{y}^{(-i)}$
3. estimation de l'erreur $\widehat{\varepsilon}_i^{(-i)} = y_i - x_i^\top \widehat{\alpha}^{(-i)}$

Fin Pour

Faut-il calculer n modèles ? NON !

Theorem (Les résidus normalisés)

$$\widehat{\varepsilon}_i^{(-i)} = \frac{\widehat{\varepsilon}_i}{1 - H_{ii}}$$

Démonstration :

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

Lemma (L'estimateur de validation croisée)

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

Démonstration

La démonstration de ce résultat est fastidieuse. Elle est basée sur le résultat suivant :

Theorem

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

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

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 $A = X^\top X$.

Pour plus de détails voir par exemple R. Christiansen "Plane answers to complex questions : the theory of linear models" Springer, 2002, p 360

l'erreur de validation croisée

on en déduit

Definition (l'erreur de validation croisée)

$$err_{VC} = \sum_{i=1}^{n} \left(\widehat{\varepsilon}_i^{(-i)} \right)^2$$

Cette quantité est notée PRESS dans la littérature anglo saxonne (*Predicted REsidual Sum of Square*)

err_{VC} se calcule pratiquement à l'aide de la formule suivante :

$$err_{VC} = \sum_{i=1}^{n} \frac{\widehat{\varepsilon}_i^2}{(1 - H_{ii})^2}$$

La divergence d'un individu

on peut normaliser les résidu de validation croisée par leur écart type.

Definition (résidu studentisés t_i)

$$t_i = \frac{\widehat{\varepsilon}_i^{(-i)}}{\sqrt{\widehat{V(\widehat{\varepsilon}_i^{(-i)})}}}$$

une interprétation des t_i est donnée par le résultat suivant :

Lemma (résidus, leviers et résidus studentisés)

$$t_i = \frac{\widehat{\varepsilon}_i}{s^{(-i)}\sqrt{1-H_{ii}}}$$

avec $(s^{(-i)})^2 = \frac{1}{n-3}\sum_{j=1,j\neq i}^{n}\left(y_j - (a^{(-i)}x_j + b^{(-i)})\right)^2$

Démonstration : puisque $\widehat{\varepsilon}_i^{(-i)} = \frac{\widehat{\varepsilon}_i}{1-H_{ii}}$ et $V(\widehat{\varepsilon}_i) = \sigma^2(1-H_{ii})$ on a $V(\widehat{\varepsilon}_i^{(-i)}) = \frac{V(\widehat{\varepsilon}_i)}{(1-H_{ii})^2} = \frac{\sigma^2}{1-H_{ii}}$, et donc :

$$t_i = \frac{\widehat{\varepsilon}_i^{(-i)}}{\widehat{V(\widehat{\varepsilon}_i^{(-i)})}} = \frac{\frac{\widehat{\varepsilon}_i}{1-H_{ii}}}{\frac{s^{(-i)}}{\sqrt{1-H_{ii}}}} = \frac{\widehat{\varepsilon}_i}{s^{(-i)}\sqrt{1-H_{ii}}}$$

Calcul pratique des résidus studentisés

$$SCE^{(-i)} = SCE - \frac{\widehat{\varepsilon}_i^2}{1 - H_{ii}} \qquad r_i^2 = \frac{\widehat{\varepsilon}_i^2}{s^2\,(1 - H_{ii})}$$

et

$$s^2 = \frac{1}{n-p} SCE \qquad (s^{(-i)})^2 = \frac{1}{n-p-1} SCE^{(-i)}$$

On a alors :

Lemma (Calcul pratique des résidu studentisés t_i)

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}$$

Démonstration :

$$\begin{aligned} t_i &= \frac{\widehat{\varepsilon}_i}{s^{(-i)}\sqrt{1 - H_{ii}}} \\ &= r_i \frac{s}{s^{(-i)}} \\ &= r_i \sqrt{\frac{n-p-1}{n-p} \frac{SCE}{SCE^{(-i)}}} \\ &= r_i \sqrt{\frac{n-p-1}{n-p} \frac{1}{1 - r_i^2/(n-p)}} \quad = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2} \end{aligned}$$

Contributions (où distance de Cook)

- Distance entre $\widehat{\alpha}^{(-i)}$ et $\widehat{\alpha}$
- Mauvaise idée : $\|\widehat{\alpha}^{(-i)} - \widehat{\alpha}\|^2$
- Bonne idée : normaliser les variables

Definition (distance de Mahalanobis))

Soient $\mathbf{x}_1$ et $\mathbf{x}_2$ deux réalisation d'une variable aléatoire gaussiènne multidimensionelle (dimension p), ayant pour espérance le vecteur μ et pour matrice de variance covariance Σ. On appelle distance de Mahalanobis entre $\mathbf{x}_1$ et $\mathbf{x}_2$

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

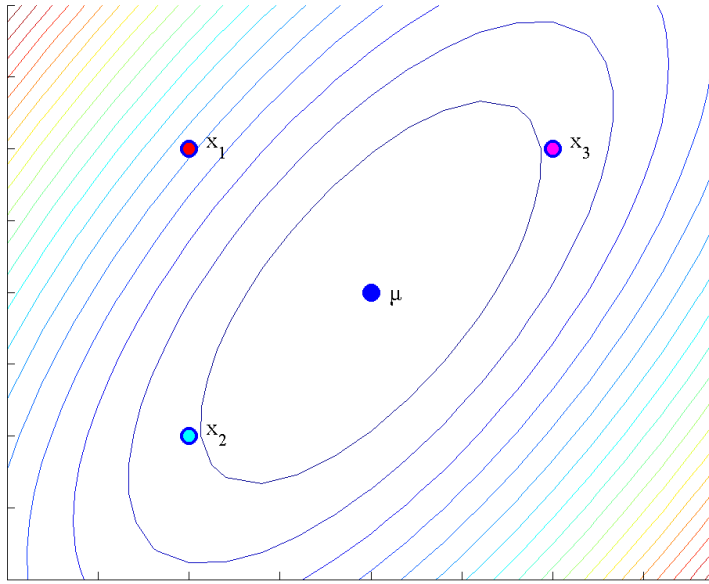


Figure: Lignes d'isocontour de la loi normale $\mathcal{N}(\mu, \Sigma)$

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = 2$$
$$\|\mathbf{x}_1 - \mathbf{x}_3\| = 3$$
$$d_M(\mathbf{x}_1, \mathbf{x}_2) = 4$$
$$d_M(\mathbf{x}_1, \mathbf{x}_3) = 2$$

Contributions (où distance de Cook)

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

$$
\begin{aligned}
\mathbb{E}(\widehat{\alpha}) &= \alpha \qquad \text{car } \mathbb{E}(\varepsilon) = 0 \\
V(\widehat{\alpha}) &= V((X^\top X)^{-1} X^\top \varepsilon) \\
&= (X^\top X)^{-1} X^\top V(\varepsilon) X (X^\top X)^{-1} \qquad \text{car } V(\varepsilon) = \sigma^2 I \\
&= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \qquad = \sigma^2 (X^\top X)^{-1}
\end{aligned}
$$

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

Definition (Contributions (où distance de Cook))

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

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

où $\widehat{\alpha}^{(-i)}$ est le vecteur des coefficients obtenu sans l'exemple $(\mathbf{x}_i, y_i)$, p le nombre de colonnes de la matrice X et $s^2 = \frac{1}{n-p} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$

Contributions (où distance de Cook)

On peut montrer que les contributions se réécrivent :

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

Soit dans le cas de la régression simple (p=2) $c_i = \frac{\sum_{j=1}^{n}(z_j^{(-i)} - z_j)^2}{p\, s^2}$

Pour le calcul pratique on montre que :

Theorem (Calcul pratique de la distance de Cook)

$$c_i = \frac{H_{ii}}{p(1-H_{ii})^2}\,\frac{\widehat{\varepsilon}_i^2}{s^2}$$

Démonstration : puisque $\widehat{\alpha}^{(-i)} = \widehat{\alpha} - \frac{(X^\top X)^{-1}\mathbf{x}_i^\top \widehat{\varepsilon}_i}{1-H_{ii}}$ on a

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

$$= \frac{\widehat{\varepsilon}_i^2}{(1-H_{ii})^2}\mathbf{x}_i(X^\top X)^{-1}\mathbf{x}_i^\top = \frac{\widehat{\varepsilon}_i^2}{(1-H_{ii})^2} H_{ii}$$

Certains auteur préconisent de se méfier d'une contribution supérieure à un.

Contributions DFFITS

Il existe d'autres mesures de contributions comme les DFFITS

Definition (Contributions DFFITS)

$$DFFITS_i = \frac{z_i^{(-i)} - z_i}{s^{(-i)}\sqrt{H_{ii}}}$$

$$DFFITS_i = t_i \sqrt{\frac{H_{ii}}{1 - H_{ii}}}$$

limite des points suspects

$$2\sqrt{\frac{p}{n}}$$

Belsley, David A.; Edwin Kuh, Roy E. Welsch (c1980). Regression diagnostics : identifying influential data and sources of collinearity, Wiley series in probability and mathematical statistics. New York: John Wiley & Sons. ISBN 0471058564.

Tableau de résultats de la régression

$$\boxed{\widehat{a} = (X^\top X)^{-1} X \mathbf{y}}$$

$$\widehat{\varepsilon}_i = y_i - \mathbf{x}_i \widehat{a} \qquad s^2 = \frac{1}{n-p} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 \qquad H = X(X^\top X)^{-1} X$$

$$r_i = \frac{\widehat{\varepsilon}_i}{s\sqrt{1 - H_{ii}}} \qquad \widehat{\varepsilon}_i^{(-i)} = \frac{\widehat{\varepsilon}_i}{1 - H_{ii}} \qquad t_i = \frac{\widehat{\varepsilon}_i}{s^{(-i)}\sqrt{1 - H_{ii}}}$$

résidus standardisés — résidus de validation croisée — résidus studentisés

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

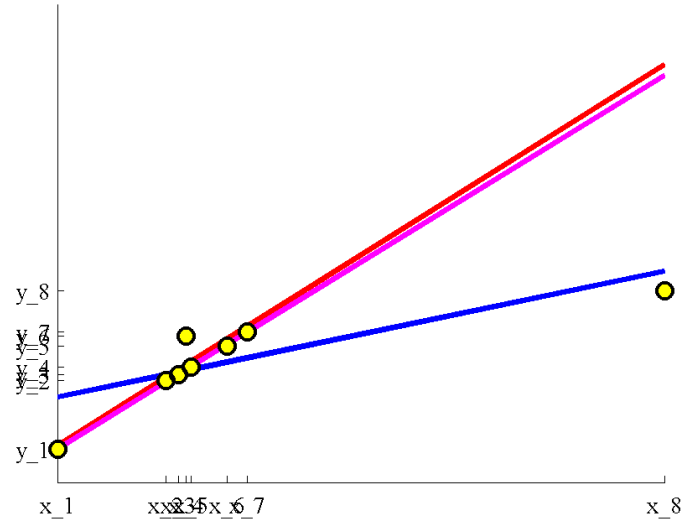
variable explicatives	variables à expliquer	résidus erreurs	résidus de VC	contributions
X	y_i	$\widehat{\varepsilon}$	$\widehat{\varepsilon}^{(-i)}$	c

Exemple

x	y	e	h	r	ei	t	c	DF
-2.00	-2.94	-3.28	0.27	-2.06	-4.50	-3.46	0.78	-2.11
0.14	1.39	-0.36	0.15	-0.21	-0.42	-0.19	0.00	-0.08
0.39	1.74	-0.17	0.14	-0.10	-0.20	-0.09	0.00	-0.04
0.55	4.21	2.19	0.14	1.26	2.54	1.35	0.13	0.54
0.64	2.26	0.18	0.14	0.11	0.21	0.10	0.00	0.04
1.35	3.53	0.99	0.13	0.57	1.13	0.53	0.02	0.20
1.75	4.46	1.65	0.13	0.95	1.89	0.94	0.06	0.35
10.00	7.04	-1.20	0.92	-2.21	-14.15	-4.64	26.31	-15.24

Tableau de régression SANS la mesure aberrante 8

x	y	e	h	r	ei	t	c	DF
-2.00	-2.94	-0.25	0.81	-0.65	-1.33	-0.61	0.92	-1.27
0.14	1.39	-0.18	0.15	-0.23	-0.22	-0.20	0.00	-0.09
0.39	1.74	-0.33	0.14	-0.40	-0.39	-0.37	0.01	-0.15
0.55	4.21	1.83	0.15	2.23	2.14	23.56	0.42	9.71
0.64	2.26	-0.30	0.15	-0.36	-0.35	-0.33	0.01	-0.14
1.35	3.53	-0.45	0.25	-0.58	-0.60	-0.54	0.06	-0.31
1.75	4.46	-0.31	0.35	-0.44	-0.48	-0.40	0.05	-0.29



régression avec tous les points (bleue)
régression sans le point x_8, y_8 (rouge)
et sans le point x_4, y_4 (magenta)

- x_1, y_1 : Fort levier – Faible erreur - Faible contribution
- x_4, y_4 : Faible levier – Forte erreur - Contribution moyenne
- x_8, y_8 : Fort levier – erreur moyenne - Forte contribution

Résumons nous

Etant données des observations $(\mathbf{x}_i, y_i), i = 1, n$

1. construire la matrice des données $X = [\mathbf{e}\ \ \mathbf{x}]$
2. calculer les coefficients : $\widehat{a} = (X^\top X)^{-1} X\mathbf{y}$
3. calculer les valeurs prédites : $\mathbf{z} = X\widehat{\alpha}$
4. calculer le R^2 : $R^2 = \frac{(z-\bar{y})^\top (z-\bar{y})}{(y-\bar{y})^\top (y-\bar{y})}$
   Si R^2 est trop petit la régression n'a pas de sens
5. calculer les résidus : $\widehat{\varepsilon} = \mathbf{y} - \mathbf{z}$
   si les résidus présentent une structure transformer les données et reprendre à 1.
6. calculer la variance estimée : $s^2 = \frac{\widehat{\varepsilon}^\top \widehat{\varepsilon}}{n-1-p}$
7. calculer les contributions : $\mathbf{c} = \frac{\mathbf{h}}{p(1-\mathbf{h})^2}\ \frac{\widehat{\varepsilon}^2}{s^2}$
   si la contribution d'un point est supérieure à $\frac{4}{n}$, examiner le point correspondant et éventuellement l'éliminer

Résumons nous : le code qui va bien

1. $X = [\mathbf{e}\ \mathbf{x}]$
2. $\widehat{a} = (X^\top X)^{-1} X\mathbf{y}$
3. $\mathbf{z} = X\widehat{\alpha}$
4. $\widehat{\varepsilon} = \mathbf{y} - \mathbf{z}$
5. $R^2 = \frac{(z-\bar{y})^\top (z-\bar{y})}{(y-\bar{y})^\top (y-\bar{y})}$
6. $H = X(X^\top X)^{-1} X$
7. $h = diag(H)$
8. $s^2 = \frac{\widehat{\varepsilon}^\top \widehat{\varepsilon}}{n-p}$
9. $\mathbf{c} = \frac{\mathbf{h}}{p(1-\mathbf{h})^2} \frac{\widehat{\varepsilon}^2}{s^2}$

```
a = (X'*X)\(X'*y)
z = X*a
r = y-z

SCM = sum((z-mean(y)).^2);
SCT = sum((y-mean(y)).^2);
R2 = SCM/SCT;
H = X*(X'*X)^(-1)*X';
h = diag(H);

s2_r = (1/(n-p))*r'*r;
t = r./((s2_r*(1-h)).^(1/2));
c = h./(p*(1-h).^2).*r.^2/s2_r;

disp('r h t c levier aberrant influent')
[r h t c   h>.5   t>2*p/n   c>4/n]
```

Plan

La pertinence des variables

$$y = a_0 + a_1 x_1 + \cdots + a_j x_j + \cdots + a_{p-1} x_{p-1} + \varepsilon$$

- et si en fait pour la variable $\mathbf{x}_j$: $a_j = 0$?
- sélection de variables
  - approche globale
  - forward
  - backward
  - stagewise

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	$y = a_0 + a_1 x_1$
1		x2			0.666	
1			x3		0.286	$\cdots$
1				x4	0.675	$y = a_0 + a_4 x_4$
1	x1	x2			0.979	$y = a_0 + a_1 x_1 + a_2 x_2$
1	x1		x3		0.548	
1	x1			x4	0.972	$\cdots$
1		x2	x3		0.847	$y = a_0 + a_3 x_3 + a_4 x_4$
1		x2		x4	0.680	
1			x3	x4	0.935	$\cdots$
1	x1	x2	x3		0.982	$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$
1	x1	x2		x4	0.982	
1		x2	x3	x4	0.973	$\cdots$
1	x1	x2	x3	x4	0.982	$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4$

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...

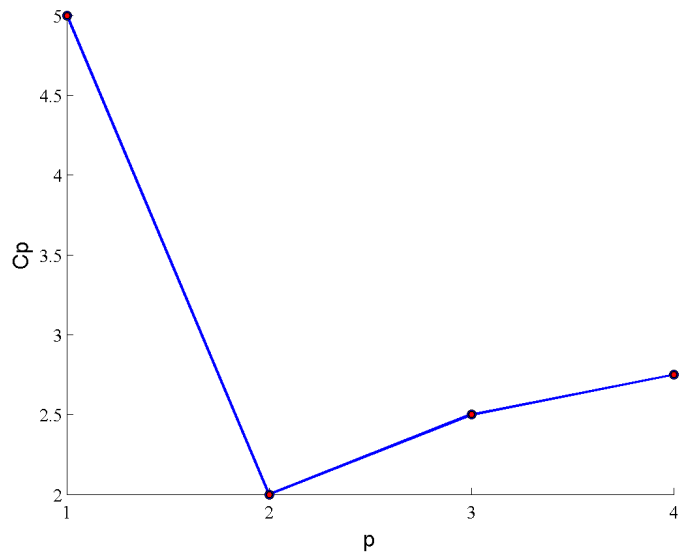
Mesure de la qualité d'un modèle : le Cp de Mallows

Definition (le Cp de Mallows)

$$Cp = \frac{1}{s^2} \sum_{i=1}^{n} (y_i - z_i^{(-i)})^2 - n + 2p$$

avec $s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - z_i)^2$ la variance estimée sur le modèle complet et p le nombre de colonnes de la matrice $X^{(-i)}$

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



Mallows, C.L., (1973) Some comments on Cp, Technometrics, 15, 661-675.

Sélection systématique des variables

1	x1	x2	x3	x4	R^2	err_VC	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
- backward selection
- stepwise regression

Conclusion

- diagnostic du modèle
  - R^2
- transformation des variables
  - Examen des résidus
- diagnostic des individus
  - examen des résidus
  - examen des contribution (distances de Cook)
- diagnostic des variables
  - calcul du Cp de Mallows
- et au dela
  - utiliser des tests statistiques
  - changer de critère : $\min_a \sum_i^n |y_i - \mathbf{x}_i a|$

Repéres bibliographiques

- Applied linear regression Par Sanford Weisberg (http://www.stat.umn.edu/alr/)
- Cook, R. Dennis (Feb 1977). "Detection of Influential Observations in Linear Regression". Technometrics (American Statistical Association) 19 (1): 15–18.
- R. Christiansen "Plane answers to complex questions : the theory of linear models" Springer, 2002
- Mallows, C.L., (1973) Some comments on Cp, Technometrics, 15, 661-675.