Régression linéaire

Plan

- Objectifs de la régression linéaire
- Lien avec la corrélation
- Régression à 2 variables ; à « p » variables
- Équations normales (sous forme matricielle);
- Somme des carrés
- Tests
- Résidus
- Amélioration d'un modèle
- Exemples

Objectifs de la régression linéaire

-Décrire le lien linéaire existant entre des variables contrôlées (X) et une variable réponse (Y)

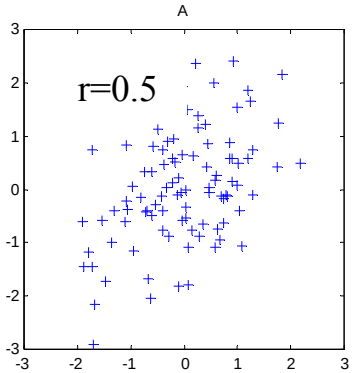
-Prédire la réponse Y en fonction de valeurs contrôlées X

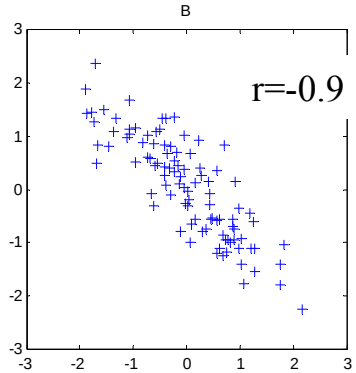
-Par extension, aussi couvrir le cas où X est une v. aléatoire non contrôlée

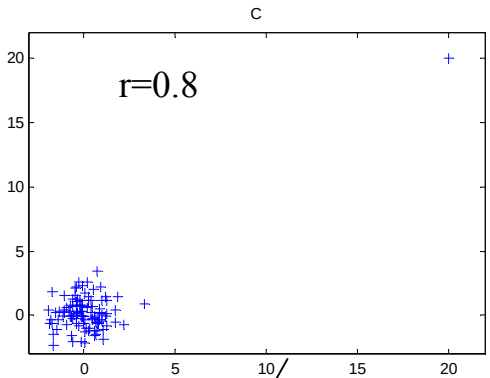
Lien avec la corrélation

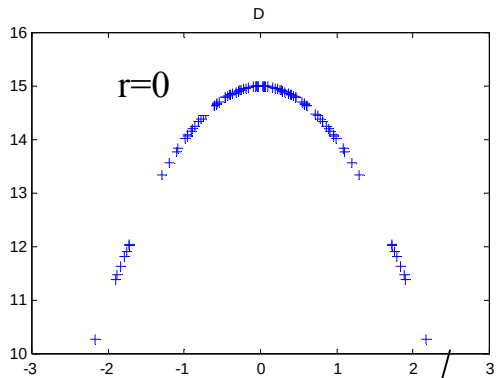
Corrélation : mesure l'intensité du lien linéaire => Est-ce que les points sont plus ou moins sur une droite dans un diagramme Y vs X ?

Régression : fournit l'équation de la droite s'ajustant le mieux (dans un certain sens) aux points









Forte corrélation

⇏

Lien fort

Absence de corrélation

⇏

Absence de lien X et Y

Équations normales sous-forme matricielle

Le modèle : $Y_i = b*X_i + e_i$ => $Y=Xb+e$

Le modèle : $Y_i = b_0 + b_1*X_i + e_i$ => $Y=Xb+e$

Le modèle : $Y_i = b_0 + b_1*X_{i,1} + b_2*X_{i,2} + \ldots + b_{p-1}*X_{i,p-1} + e_i$ => $Y=Xb+e$

=> Tous les modèles ont la même forme en écriture matricielle !

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & X_{1p} \\ 1 & X_{21} & X_{22} & . & . & X_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & X_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$Y = Xb + e$$

$$\hat{Y} = Xb$$

$$\Rightarrow e = Y - \hat{Y} = Y - Xb$$

On veut minimiser la somme des carrés des erreurs (SCE)

SCE = e'e = (Y-Xb)'(Y-Xb)

$$SCE = Y'Y - Y'Xb - b'X'Y + b'X'Xb$$

$$\frac{\delta SCE}{\delta b} = 0 = (X'X)b - X'Y$$

$$\boxed{b = (X'X)^{-1} X'Y}$$

Exemple (modèle avec constante)

On a 2 observations : $(X_1, Y_1) = (1, 1.5), (X_2, Y_2) = (2, 2)$

$$Y = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 3.5 \\ 5.5 \end{bmatrix} \qquad b = (X'X)^{-1}X'Y = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

$$\hat{Y} = Xb = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}; e = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

On retrouve l'équation de la droite passant par les 2 points

Exemple (modèle sans constante)

On a 2 observations : $(X_1, Y_1) = (1, 1.5), (X_2, Y_2) = (2, 2)$

$$Y = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

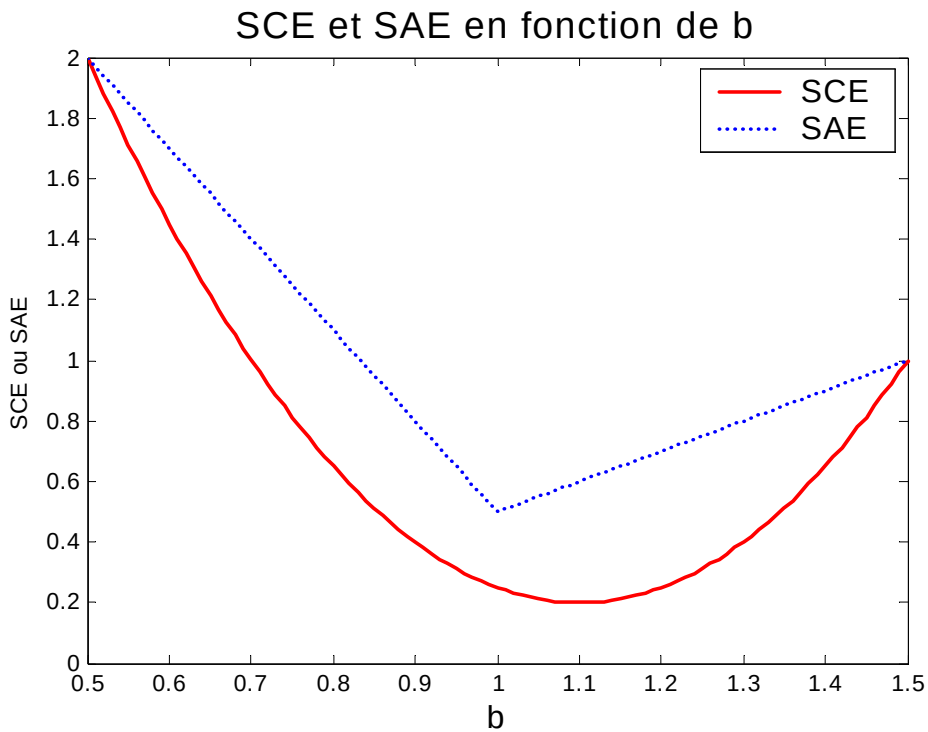
$$X'X = 5 \qquad (X'X)^{-1} = 1/5$$

$$X'Y = 5.5 \qquad b = (X'X)^{-1}X'Y = 5.5/5 = 1.1$$

$$\hat{Y} = Xb = \begin{bmatrix} 1.1 \\ 2.2 \end{bmatrix}; e = \begin{bmatrix} 0.4 \\ -0.2 \end{bmatrix}$$

La somme des résidus ne donne pas 0 car modèle sans constante

Modèle sans constante



Partition en somme des carrés

(modèle avec constante)

Nom	Sigle	Définition		d.l.	Remarques
S.c. totale	SCT	$Y'Y$	Σy_i^2	n	
S.c. totale corrigée pour la moyenne	SCT_m	$(Y-Y_m)'(Y-Y_m)$	$\Sigma(y_i-y_m)^2$	n-1	
S.c.de la moyenne	SCM	$Y_m'Y_m$	ny_m^2	1	$SCT=SCT_m+SCM$ $SCM \perp SCT_m$
S.c. de la régression	SCR	$Y_p'Y_p$	Σy_{pi}^2	p+1	
S.c. de la régression sans la moyenne	SCR_m	$(Y_p-Y_m)'(Y_p-Y_m)$	$\Sigma(y_{pi}-y_m)^2$	p	$SCR=SCR_m+SCM$ $SCM \perp SCR_m$
S.c. erreur	SCE	$e'e$	Σe_i^2	n-(p+1)	$SCT=SCR+SCE$
		$(Y-Y_p)'(Y-Y_p)$			$SCT_m=SCR_m+SCE$
					$SCE \perp SCR$
					$SCE \perp SCR_m$

Autre forme pour les sommes de carrés

Somme des carrés	Forme classique	Forme « idempotente »
SCT	$Y'Y$	$Y'IY$
SCM	$Y_m'Y_m$	$Y'(11'/n)Y$
SCT_m	$(Y-Y_m)'(Y-Y_m)$	$Y'(I -11'/n)Y$
SCR	$Y_p'Y_p$	$Y'MY$
SCR_m	$(Y_p-Y_m)'(Y_p-Y_m)$	$Y'(M-11'/n)Y$
SCE	$e'e$	$Y'(I-M)Y$

$M=X(X'X)^{-1}X'$

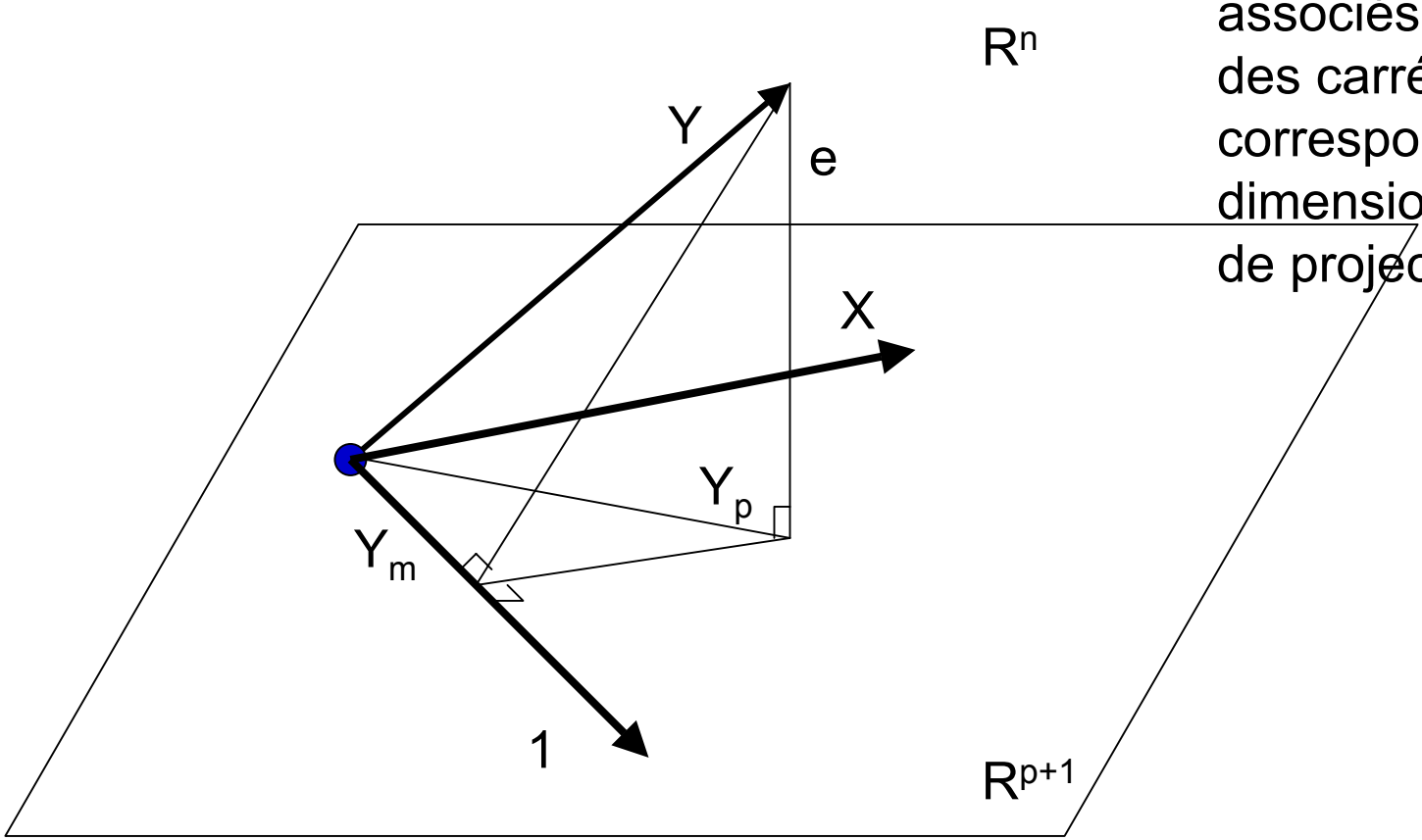
Matrice idempotente : *Matrice de projection*

-Toutes les quantités (Y, Y_m, Y_p, e) sont donc des projections du vecteur Y

-Toutes les sommes des carrés sont les longueurs au carré de ces projections

Géométrie des moindres carrés

(modèle avec constante)



Les degrés de liberté associés aux sommes des carrés correspondent à la dimension de l'espace de projection

Tests en régression

Toutes les sommes de carrés: forme Y'AY où A est symétrique et idempotente

$Y=X\beta + \varepsilon$ avec $\varepsilon \sim N(0,\sigma^2)$

$\Rightarrow Y \sim N(X\beta,\sigma^2)$

$\Rightarrow Y/\sigma \sim N(X\beta/\sigma,1)$

$\Rightarrow (Y/\sigma)^2 \sim \chi^2_{1,\delta}$ (Si $Z \sim N(0,1)$ alors $Z^2 \sim \chi^2_1$)

$\Rightarrow Y' I Y/\sigma^2 \sim \chi^2_{n,\delta}$ (Somme de « n » χ^2_1 ind est χ^2_n)

Par extension, $Y'AY/\sigma^2 \sim \chi^2_{rang(A),\delta(A)}$

σ^2 est inconnue => prendre le rapport de 2 sommes des carrés

La distribution de $(\chi^2_{n1}/n1) / (\chi^2_{n2}/n2)$ est une loi $F_{n1,n2}$ si:

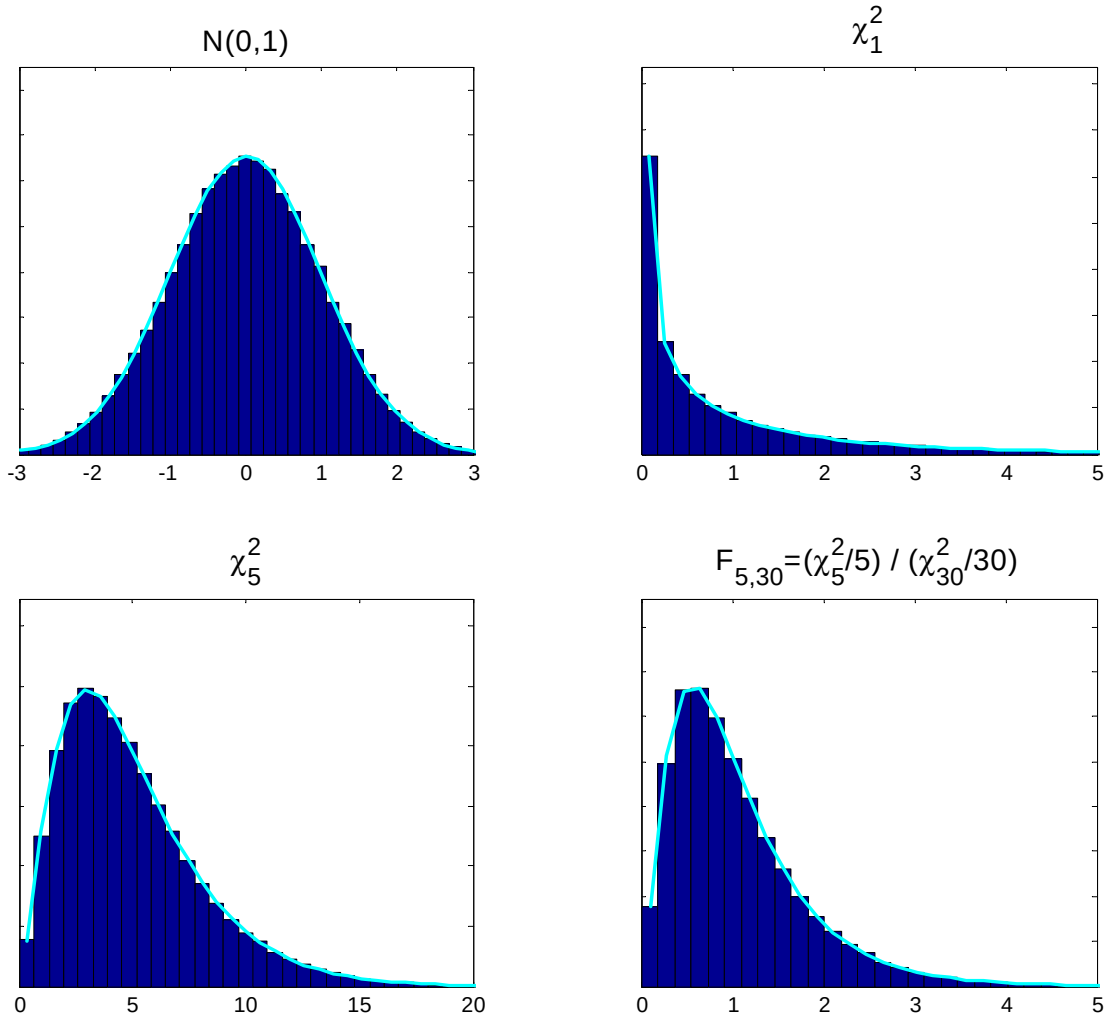
- les variables sont indépendantes

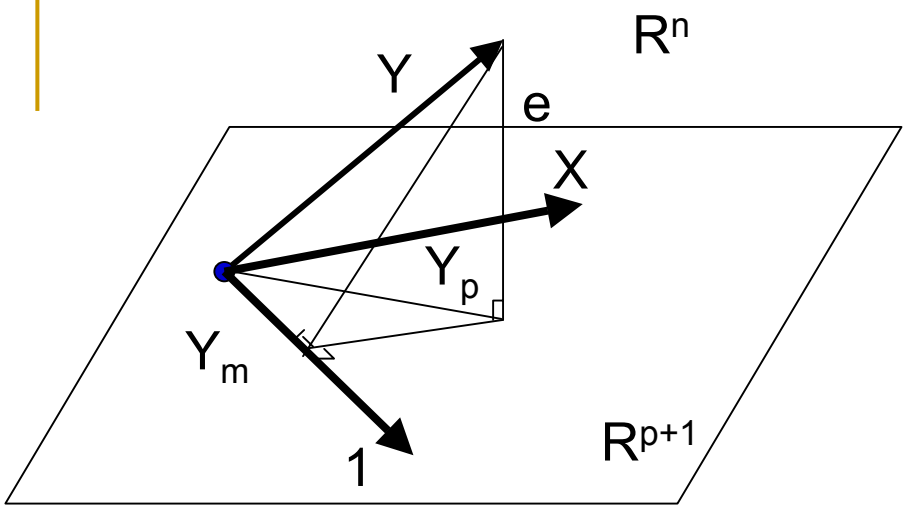
(=> les vecteurs associés sont orthogonaux => AB=0, où A et B sont les matrices idempotentes associées aux projections)

- le paramètre de non-centralité $\delta = 0$ pour chaque χ^2

C'est le cas pour SCE. Pour les autres SC, ce sera le cas sous l'hypothèse H_0 du test

(Digression : quelques exemples de distributions)





$e \perp 1, X, Y_m, Y_p, (Y_p - Y_m)$

$e \sim N(0, (I-M)\,\sigma^2) \Rightarrow SCE/\sigma^2 \sim \chi^2_{n-(p+1)}$

$Y_p \sim N(X\beta, M\,\sigma^2) \Rightarrow SCR/\sigma^2 \sim \chi^2_{(p+1)}$ avec $\delta = Y'MY/\sigma^2 = \beta'X'X\beta/\sigma^2$

Sous $H_0 : \beta=0 \Rightarrow \delta = 0 \Rightarrow$

$$\frac{SCR}{p+1} \Big/ \frac{SCE}{n-(p+1)} \sim F_{p+1,n-(p+1)}$$

Test peu utile
Pourquoi ?

$e \perp (Y_p - Y_m)$

Sous $H_0 : \beta=0$ (sauf β_0) => $E[(Y_p - Y_m)]=0$ => $\delta = 0$

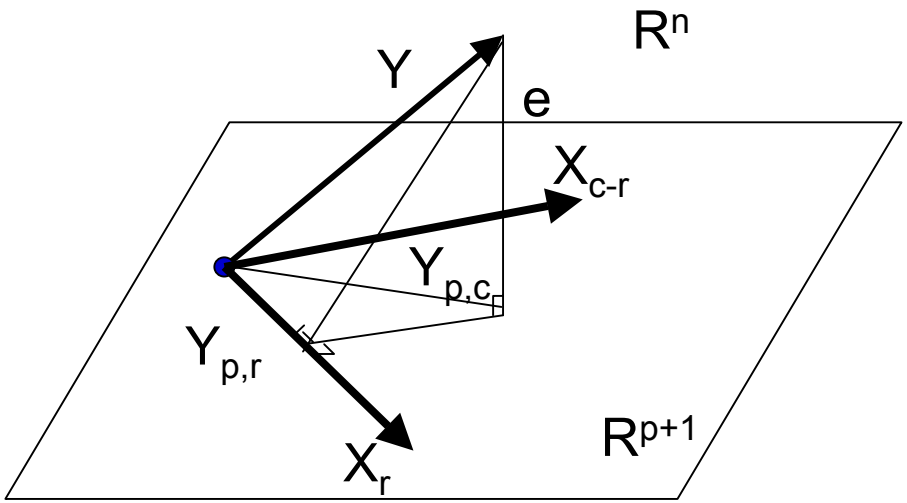
$$\frac{SCR_m}{p} \Big/ \frac{SCE}{n-(p+1)} \sim F_{p,n-(p+1)}$$

Permet de tester si la régression est significative

Si le modèle est sans constante

$$\frac{SCR}{p} \Big/ \frac{SCE}{n-p} \sim F_{p,n-p}$$

Généralisation des tests : modèle complet vs modèle réduit



$e \perp (Y_{p,c} - Y_{p,r})$

Sous H_0 : Les paramètres ajoutés pour passer du modèle réduit au modèle complet sont tous nuls

$\Rightarrow E[(Y_{p,c} - Y_{p,r})] = 0 \Rightarrow \delta = 0$

$$\frac{SCR_c - SCR_r}{(c-r)} \Big/ \frac{SCE}{n-c} \sim F_{c-r,n-c}$$

où c: nombre de paramètres dans le modèle complet (incluant la constante s'il y a lieu)

r : nombre de paramètres du modèle réduit

Exemples de tests

On a 10 observations d'une variable Y et de 3 variables X.

Modèle : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ avec $\varepsilon \sim N(0, \sigma^2)$

On trouve: b0 = -7.45, b1=2.06, b2=2.15, b3=-1.57

On obtient la table « anova » suivante:

	SC	d.liberté
SCE	1898	6
SCR_m	58771	3

Les variables X expliquent-elles Y?

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$F_{calculé} = (58771 / 3) / (1898 / 6) = 61.9$

$> F_{table} = F_{3,6;\ 0.05} = 4.76$

=> La régression est significative

Lorsque l'on effectue la régression seulement avec X_1 et X_2, on obtient : SCE= 14249 (d.l.=7) et SCR_m= 46420 (d.l.=2)

Vaut-il la peine d'inclure la variable X_3 ?

$H_0 : \beta_3 = 0$

$F_{calculé}$=[(58771-46420) / (3-2)] / (1898 / 6) = 39.04

➢ $F_{table} = F_{1,6;\ 0.05} = 5.99$

=> Oui, X_3 améliore significativement la régression.

Noter que c'est SCE_c et non SCE_r qui apparaît au dénominateur

Peut-on considérer que $\beta_1 = \beta_2$?

On définit les modèles réduit et complet suivants:

Modèle réduit :

$$Y = \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 X_3 + \varepsilon$$

Modèle complet :

$$Y = \beta_0 + \beta_1 (X_1 + X_2) + (\beta_2 - \beta_1) X_2 + \beta_3 X_3 + \varepsilon$$

On obtient : $SCE_r = 1916.5$ (d.l.=7), $SCE_c = 1898$ (d.l.=6)

$H_0 : \beta_2 - \beta_1 = 0$

$F_{calculé} = [(1916.5-1898) / (7-6)] / (1898 / 6) = 0.06$

$< F_{table} = F_{1,6;\ 0.05} = 5.99$

$\Rightarrow$ On ne peut rejeter l'hypothèse que $\beta_1 = \beta_2$

On dispose d'un modèle théorique pour ce problème:

$$Y = 20 + 2X_1 + 2.3X_2 - 2X_3 + \varepsilon$$

Nos données s'écartent-elles significativement de ce modèle ?

Méthode: on calcule le $Y_{théorique} = 20 + 2X_1 + 2.3X_2 - 2X_3$

On calcule $Y_{observé} - Y_{théorique} = \varepsilon$

Si nos données sont compatibles avec le modèle théorique, la régression de $Y_{observé} - Y_{théorique}$ sur X_1 X_2 X_3 et « cte » devrait être non-significative

On obtient : SCR = 2334.1 (d.l.=4), SCE=1898 (d.l.=6)

$H_0 : \beta_0 = 20,\ \beta_1 = 2, \beta_2 = 2.3,\ \beta_3 = -2$

$F_{calculé} = (2334.1/4) / (1898 / 6) = 1.84$

$< F_{table} = F_{4,6;\ 0.05} = 4.53$

$\Rightarrow$Les données ne s'écartent pas significativement du modèle théorique

Parmi les 10 données 6 proviennent du laboratoire A et 4 du laboratoire B.

Y-a-t-il un « effet laboratoire » sur les données ?

Permettre un modèle différent pour chaque laboratoire et tester s'ils sont significativement différents (modèle complet vs modèle réduit)

Modèle réduit:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Modèle complet:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 I + \beta_6 I X_1 + \beta_7 I X_2 + \beta_8 I X_3 + \varepsilon$$

où I est une variable indicatrice indiquant la provenance de la donnée (0 si A, 1 si B)

On obtient SCE_r=1898 (d.l.=6) SCE_c= 390.8 (d.l.=2)

$H_0 : \beta_5 = 0, \ \beta_6 = 0, \beta_7 = 0, \beta_8 = 0$

$F_{calculé} = [(1898-390.8)/(6-2)] \ / (390.8 / 2) = 1.93 \quad < \quad F_{table} = F_{4,2;\ 0.05} = 19.2$

Les données n'indiquent pas « d'effet laboratoire »

Distributions et intervalles de confiance

Des équations normales découle:

$$Var(b) = \sigma^2 (X'X)^{-1}$$

$$Var(Y_{p,i}) = Var(x_i b) = \sigma^2 (x_i (X'X)^{-1} x_i')$$

$$Var(Y_i - Y_{p,i}) = \sigma^2 (1 + x_i (X'X)^{-1} x_i')$$

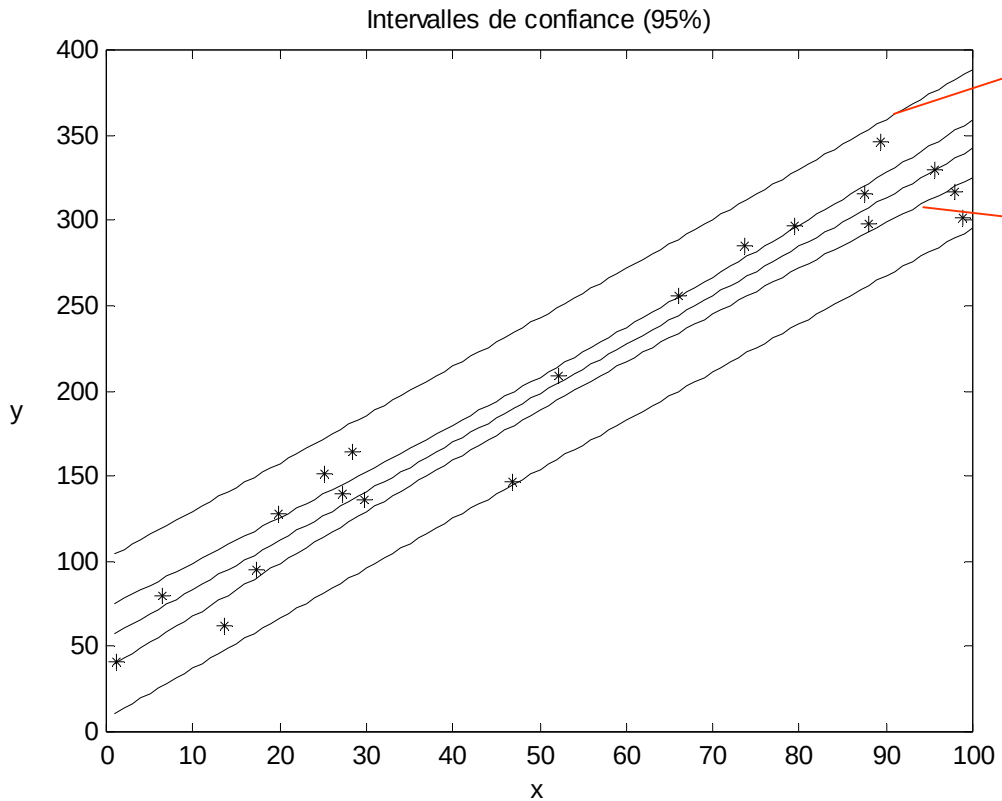
Pour construire les intervalles de confiance

σ^2 inconnue => prendre SCE / (d.l. de SCE)

=> loi normale -> Student avec d.l. de SCE

Note : la plupart des logiciels donnent l'écart-type de b (obtenu de $\sigma^2 (X'X)^{-1}$) et la statistique « t » correspondante. Cette approche est strictement équivalente au test d'ajout d'une seule variable. Si plus d'une variable montre un « t » non-significatif, on ne peut conclure que l'ajout simultané de ces variables est non-significatif. Seul le test d'ajout permet de faire cette vérification

Exemple d'intervalle de confiance



Intervalle pour $(Y_i - Y_{pi})$

Intervalle pour Y_{pi}

Coefficient de détermination

Proportion de la variation de Y expliquée par le modèle.

Modèle avec constante:

$$R^2 = SCR_m/SCT_m \qquad (0 \leq R^2 \leq 1)$$

Modèle sans constante :

$$R^2 = 1 - SCE/SCT_m \quad (R^2 \leq 1;\ \text{peut être négatif})$$

Note :

- *en ajoutant une variable R^2 ne peut pas décroître*
- *si le nombre de paramètres = n => R^2 = 1 peu importe la valeur des paramètres*

Examen des résidus

Étape *très importante* pour :

- juger de la validité d'un modèle
- identiifer des améliorations possibles
- détecter des données « particulières »

Si toutes les hypothèses sont vérifiées les résidus devraient avoir les caractéristiques suivantes:

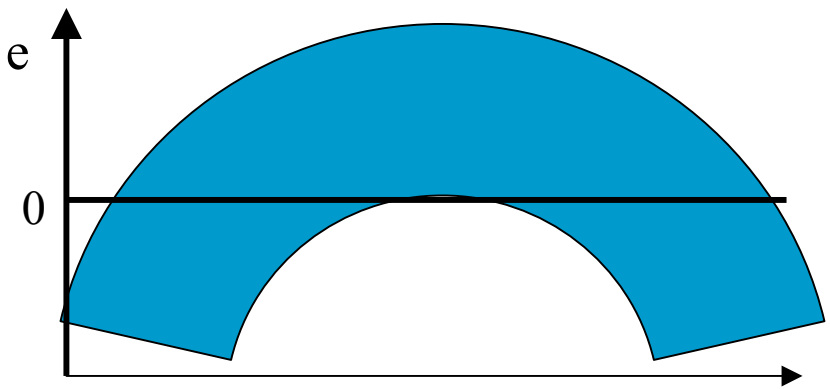
- distribution normale (globalement)
- répartition symétrique autour de $Y_{p,i}$
- absence d' autocorrélation entre les résidus lorsque triés selon X ou $Y_{p,i}$ ou toute autre variable d'intérêt (e.g. temps, provenance, etc.)

Note importante : $Cov(e,Y) = \sigma^2(I-M)$ et e'Y=SCE

Il est donc normal d'observer une corrélation entre e et Y !

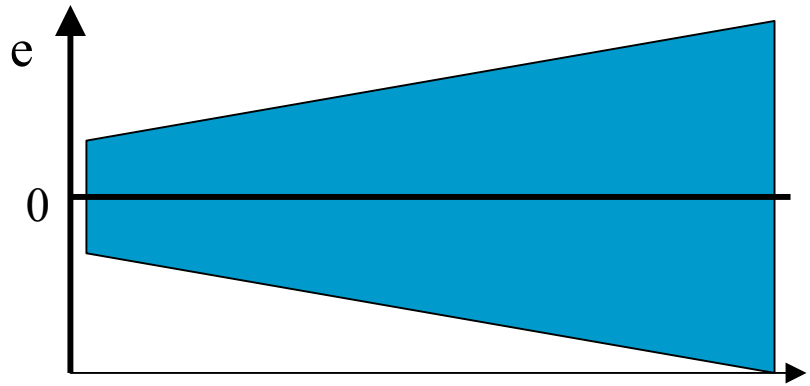
Patron attendu

X_i , Y_p, variables externes



Patron en arche, inclure des monômes (e.g. $x^{0.5}$, x^2, etc)

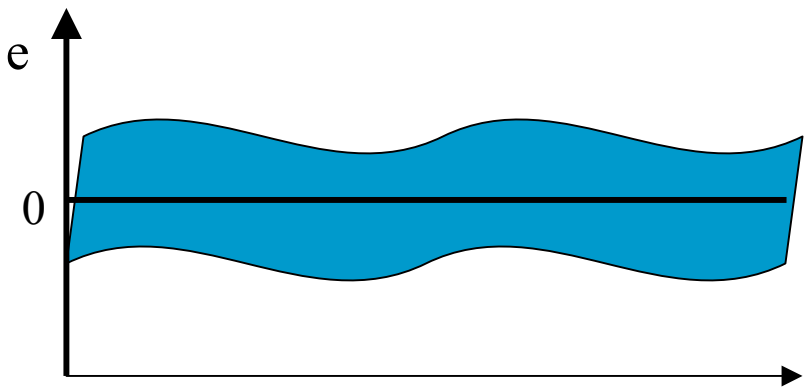
X_i , Y_p, variables externes



Patron en éventail indique hétéroscédasticité des résidus du modèle

=> log(Y) ?? log(X)

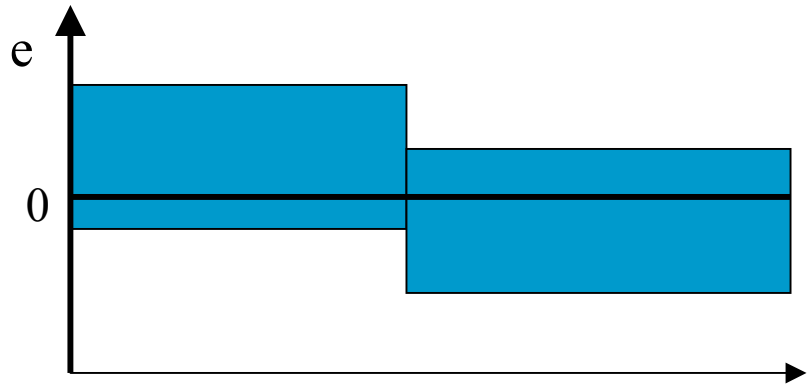
X_i , Y_p, variables externes



Patron périodique

inclure une composante temps ?

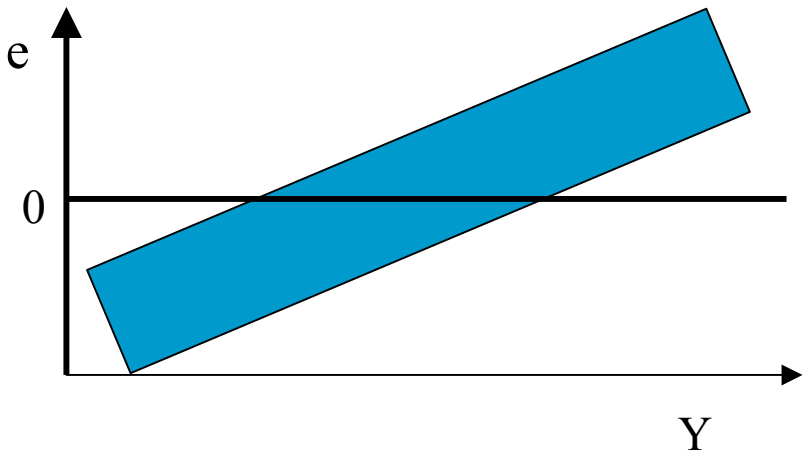
X_i , Y_p, variables externes



Patron avec différence de niveau

=> Inclure l'effet de la variable externe (variable indicatrice)

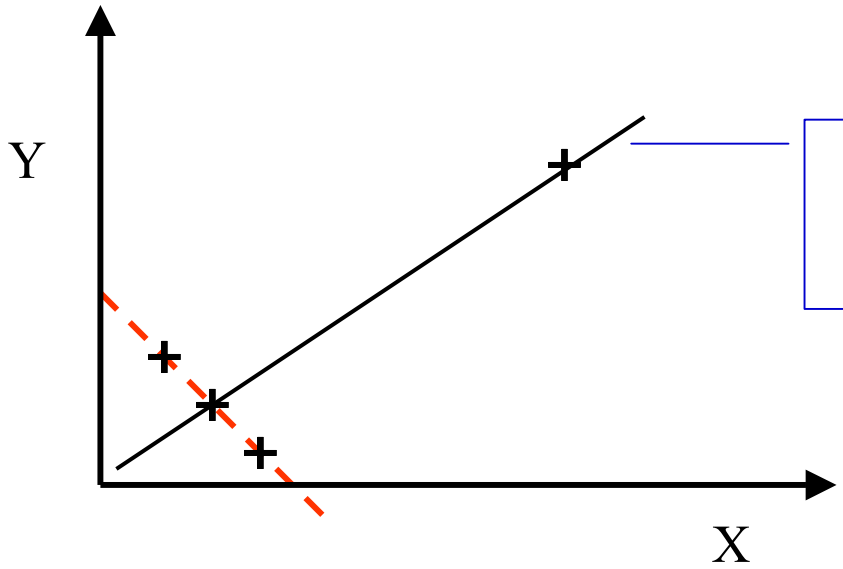
variables externes (e.g. laboratoire, échantillonneur, appareil, temps, etc.)



Patron attendu étant donné e'Y=SCE

Mauvais choix de graphe!

Influence d'une observation



La droite de régression est déterminée par cette observation. Si on l'enlève, la droite tourne de presque 90°

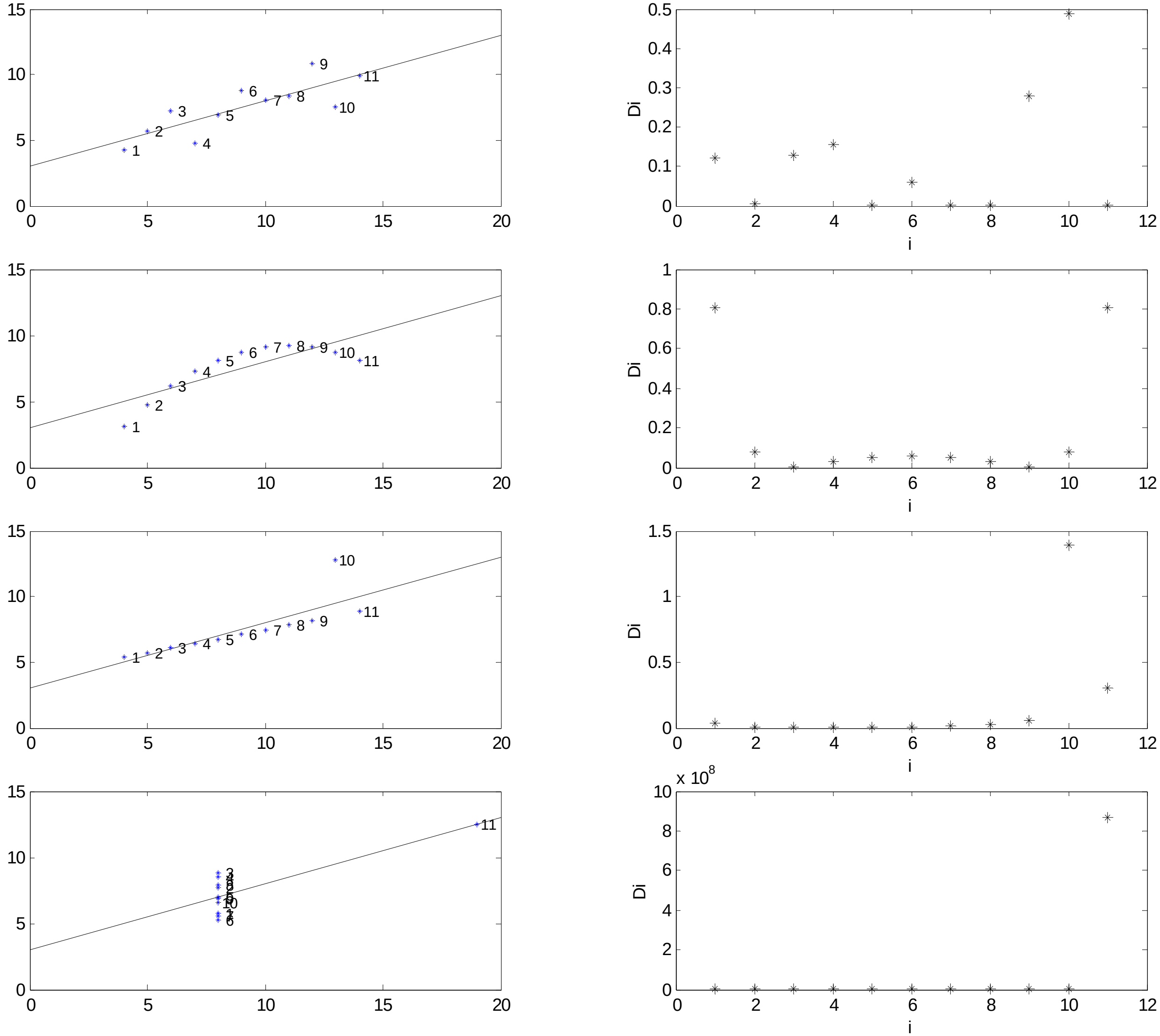
Note: régression significative, résidus ok, l'observation influente montre un résidu très faible

Pour détecter ce problème, faire la régression en enlevant 1 donnée à tour de rôle et calculer de combien la droite de régression tourne

Influence d'une observation
(distance de Cook)

$$D_i = \frac{(b_{(i)} - b)'(X'X)(b_{(i)} - b)}{(p+1)CME} = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{(p+1)CME}$$

D_i >1 => forte influence de l'observation « i »



Toutes les régressions ont même b et même R^2

Sélection de variables

Plusieurs variables plausibles pour expliquer Y

Toutes les inclure ? => modèle surajusté aux données, non-robuste, pauvres performances

$\Rightarrow$choisir les variables les plus performantes

Stratégies :

- Tous les sous-ensembles de variables : prohibitif si « p » grand (2^p-1 sous-ensembles)

- Sélection avant : ajouter une variable à la fois tant que l'ajout est significatif (voir tests)

- Élimination arrière : éliminer une variable à la fois tant que la variable éliminée n'est pas significative

- « stepwise » (pas à pas) : alterner sélection avant et élimination arrière tant que c'est possible.

Ces stratégies ne sont que des aides !

-Modèle obtenu pas nécessairement le plus « naturel »

User de bon sens et de ses connaissances : *ne pas hésiter à modifier l'ensemble de variables si physiquement plus plausible. Souvent le R^2 à peine affecté !*

Coefficients de régression normalisés

Coefficients « b » dépendent des unités

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p + e$$

On peut réécrire

$$Y - \overline{Y} = b_1 (X_1 - \overline{X}_1) \frac{s_1}{s_1} + \ldots + b_p (X_p - \overline{X}_p) \frac{s_p}{s_p} + e$$

Ou encore

$$Y - \overline{Y} = b_1^* X_1^* + \ldots + b_p^* X_p^* + e$$

Les b_i^* sont les coefficients normalisés de la régression

$$b_i^* = b_i s_i \qquad X_i^* = \frac{X_i - \overline{X}_i}{s_i}$$

Permettent de mieux juger de l'importance relative d'une variable dans la régression (meilleur numériquement)

Ex. (aspect numérique)

Données en coordonnées MTM prises sur la rive sud

Moyenne des coordonnées (2.7e05, 5e06)

X= [MTM_est, MTM_nord **1**]

Nombre de conditionnement de X'X= 1.6e23 ! => inv(X'X) matlab retourne matrice quasi-singulière

En prenant: $X_i^* = \frac{X_i - \overline{X}_i}{s_i}$

Nombre de conditionnement de X*'X*= 1.4 => inv(X*'X*) ne pose pas de problème

Multicollinéarité

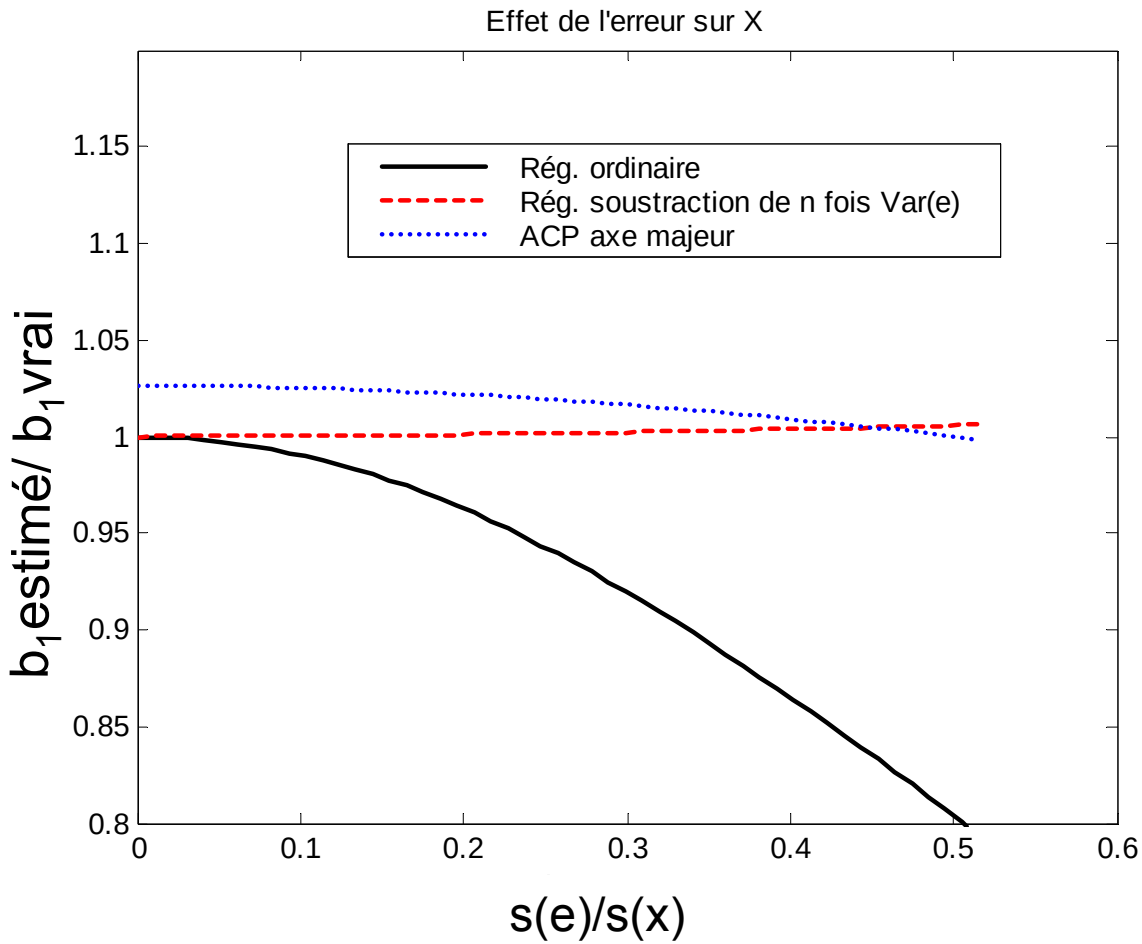
Si une des variables X est parfaitement corrélée aux autres variables X => (X'X) est singulière et les coefficients « b » ne peuvent pas être calculés

Si une des variables X est presque parfaitement corrélée aux autres variables X => (X'X) est quasi-singulière et les coefficients « b » sont instables. On peut s'en rendre compte par Var(b) qui devient élevé

(diag(Var(b)/CME) est appelé « variance inflation factor », les valeurs supérieures à 10 indiquent un problème)

Avec les procédures de sélection avant et pas à pas => pas un problème car une variable X très corrélée aux autres ne peut jamais être incluse dans la régression.

Régression quand X est une v.a.



Moindres carrés pondérés

Parfois, certains Y_i sont mesurés avec une plus grande précision que d'autres.

On peut en tenir compte en minimisant e'We au lieu de e'e

W est une matrice diagonale nxn (e.g. $1/\sigma_i^2$ sur la diagonale) qui donne plus de poids aux données précises.

Ceci même aux équations suivantes:

$$b=(X'WX)^{-1}X'WY$$

$$Var(b)=(X'WX)^{-1}\sigma^2$$

Tout le reste demeure virtuellement inchangé

Moindres carrés généralisés

On peut généraliser l'idée précédente en utilisant :

W matrice nxn = Σ^{-1} avec Σ la matrice de variances-covariances des Y_i

Ceci mène aux mêmes équations que pour le cas pondéré

Erreur pure et manque d'ajustement

Lorsque l'on peut contrôler X => plusieurs observations pour un même X

⇒Permet de décomposer le vecteur « e » en 2 composantes orthogonales:

$$e = (Y-Y_p) = \underbrace{(Y-Y_{m(x)})}_{\text{Erreur pure}} + \underbrace{(Y_{m(x)}-Y_p)}_{\text{Manque d'ajustement}}$$

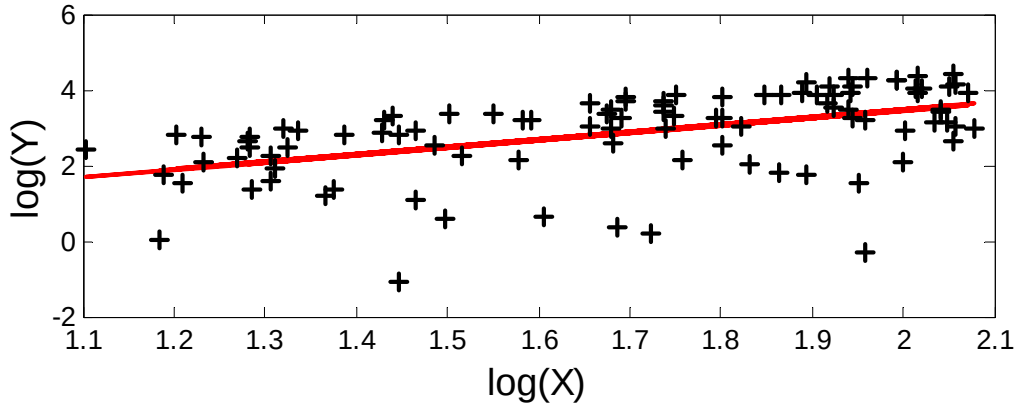
$Y_{m(x)}$ est un vecteur n x 1 avec, sur chaque ligne, la moyenne des Y pour le X correspondant

Peut tester manque d'ajustement vs erreur pure. Si significatif, modèle doit être amélioré, sinon, on procède comme d'habitude.

Transformations de Y

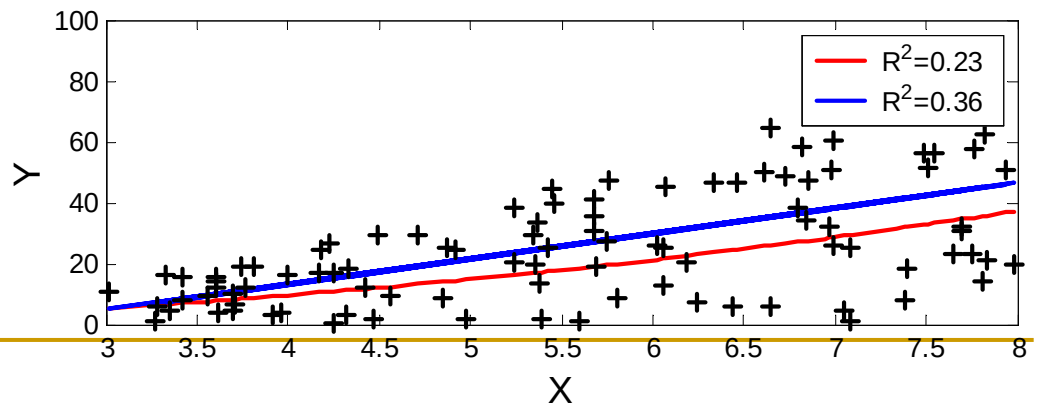
Il faut éviter autant que possible de transformer les Y

-La transformation inverse normalement introduit un biais (i.e. les « e » ne sont plus de moyenne 0
-La courbe de régression obtenue n'est pas optimale dans l'espace transformée (i.e. e'e pourrait être réduit avec le même modèle)



$\Sigma e = 6.9$!

(Moy(Y)=27)



Régression polynomiale

Un modèle du type :

$$Y_i = b_0 + b_1X + b_2X^2 + b_3X^3 + \ldots + b_pX^p + e$$

Est une fonction linéaire des paramètres inconnus b_0 , $b_1 \ldots b_p$

Cas particulier : estimation d'une dérive («Trend surface » p. 22)

Plusieurs méthodes basées sur des processus stationnaires (e.g. transformée de Fourier en traitement de signal, variogramme en géostatistique,…)

Idée: estimer la *moyenne par un polynôme de faible degré* et effectuer le traitement sur les résidus

Ex. en 2D: $m(x,y) = b_{00} + b_{10}\,x + b_{01}\,y + b_{20}\,x^2 + b_{11}\,xy + b_{02}\,y^2 + \ldots + e$

« e » devient la variable d'intérêt

Exemples d'application

Correction géométrique de photos (p. 23)

- photo A (ou carte) référence (repère u,v)
- photo B présentant distorsion (repère x,y) que l'on veut corriger

i. Identifier « n » points de contrôle sur les 2 photos (points facilement repérables (e.g. bâtiment, linéament, route, bloc particulier, etc.) et couvrant l'ensemble de la photo

ii. Effectuer deux régressions polynomiales (faible degré) :
x en fonction de (u,v)
y en fonction de (u,v)

iii. Pour chaque position (u_0,v_0) de A, calculer x_0^* et y_0^* par régression. Représenter à la position (u_0,v_0), la valeur lue sur la photo B à la position (x_0^*,y_0^*) => photo B corrigée