2. CORRÉLATION ET RÉGRESSION

2.1 Introduction

La meilleure façon de décrire la relation unissant deux variables est de construire un diagramme binaire ("scatterplot") de ces deux variables. Ce diagramme renferme toute l'information sur le comportement conjoint des deux variables. Lorsqu'un lien linéaire (pas nécessairement parfaitement linéaire) existe entre ces deux variables, on peut être intéressé à le quantifier à l'aide d'une mesure numérique unique qui permettra d'établir des comparaisons entre la force des liens linéaires unissant diverses paires de variables.

La mesure qui permet de quantifier la force de ce lien linéaire s'appelle **coefficient de corrélation (simple)**.

2.2 Coefficient de corrélation simple

On définit le coefficient de corrélation simple par:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \ \sigma_y} \qquad 2.1$$

où σ_x est l'écart-type de la variable X
et σ_{xy} est la covariance entre les variables X et Y

On se rappellera que:

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] \qquad 2.2$$

et

$$\sigma_x^2 = E[(X - \mu_x)^2] \qquad 2.3$$

μ_x et μ_y sont les moyennes des variables X et Y.

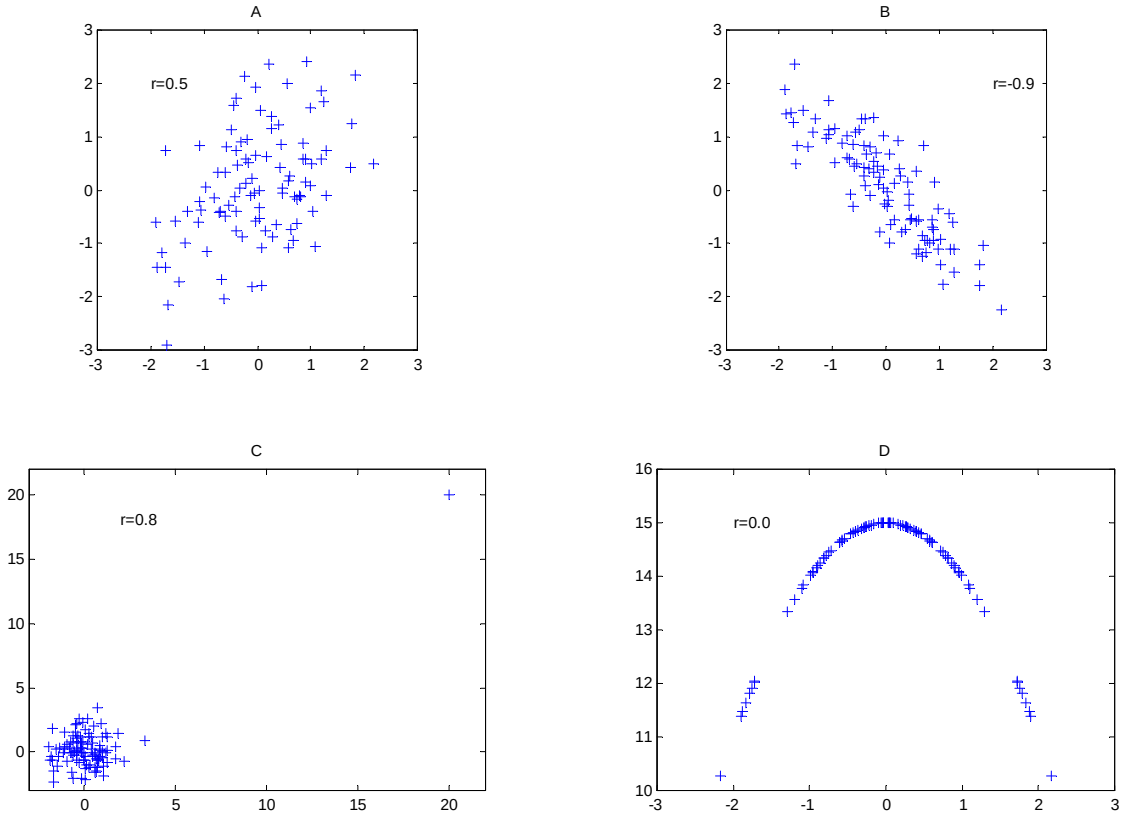
La variance mesure la dispersion (carrée) moyenne autour de la moyenne de la variable X. L'écart-type (σ) en est la racine carrée. La covariance mesure si les dispersions des deux variables autour de leurs moyennes se produisent indépendamment (covariance nulle) ou si elles sont liées (positivement ou négativement).

En fait, covariance et corrélation sont deux notions soeurs. Toutefois, alors que la covariance possède des unités et, conséquemment, varie selon le choix des unités de mesure, la corrélation, elle, est sans unité, et est donc invariable face au choix des unités de mesure.

Question 1: Comment la covariance et la corrélation sont-elles affectées par l'ajout d'une constante à la variable X? Par la multiplication par une constante? Pouvez-vous le démontrer?

Une corrélation est toujours comprise entre -1 et 1 inclusivement.

L'absence de corrélation n'implique pas l'indépendance entre les variables. Elle implique uniquement l'absence de relation linéaire entre celles-ci. Par contre, l'indépendance entre les variables implique l'absence de corrélation.



Question 2: Comment décririez-vous la corrélation observée en C? Quelle pourrait-en être la cause? Que ceci suggère-t-il?

Question 3: En D, suggérez une transformation de la variable X qui permettrait l'apparition d'une corrélation de 1.0 entre les deux variables. Que ceci vous suggère-t-il lorsque vous etudiez un jeu de données et êtes à la recherche de corrélations fortes? Concluez quant à l'utilité des diagrammes binaires.

En pratique on estime la corrélation, à partir d'un échantillon, à l'aide de:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)}} \qquad 2.4$$

qu'on peut aussi écrire:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\left(\sum_{i=1}^{n} (x_i y_i) - n\bar{x}\bar{y} \right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right)}} \qquad 2.5$$

2.3 Régression linéaire entre deux variables

Une fois constatée l'existence d'un lien linéaire entre deux variables, il peut être intéressant de chercher à décrire l'équation de la droite ayant le meilleur ajustement possible (en termes de moindres carrés) au nuage de points. Contrairement à la corrélation, le problème ici n'est pas entièrement symétrique. En régression, on doit déterminer une variable "à expliquer" et une variable "explicative", i.e., on a un modèle sous-jacent de la forme suivante

$$y_i = b_0 + b_1 x_i + e_i \qquad 2.6$$

où y_i est la ième observation de la variable à expliquer,
x_i est la ième observation de la variable explicative,
e_i est le résidu entre la droite (estimée) et la valeur réellement observée (y_i).

Dans cette équation, b_0 et b_1 représentent les paramètres (estimés) de la droite donnant le meilleur ajustement au sens des moindres carrés. Clairement, si on intervertit les rôles de x et y, il n'y a aucune raison pour que b_0 et b_1 demeurent inchangés.

On peut montrer que les coefficients b_0 et b_1 sont donnés (dans le cas de la régression de y sur x) par:

$$\mathbf{b}_0 = \bar{\mathbf{y}} - \mathbf{b}_1 \bar{\mathbf{x}}$$
$$\mathbf{b}_1 = \frac{\mathbf{s}_{xy}}{\mathbf{s}_x^2} \qquad 2.7$$

On n'a qu'à intervertir x et y dans ces équations pour obtenir les coefficients de la régression de x sur y.

Question 4: Si le coefficient de corrélation est zéro, quel sera l'angle entre les deux droites de régression? Si le coefficient de corrélation est 1, quel est l'angle entre les deux droites? Qu'arrive-t-il dans ce cas? Faites les démonstrations. Qualitativement, comment varie l'angle entre les deux droites en fonction de r_{xy}?

Si on a le modèle $y=b_0+b_1x+e$ et le modèle $x=c_0+c_1y+e$
Peut-on dire que $c_1=1/b_1$?

Remarque: A proprement parler, la droite précédente devrait être appelée droite des moindres carrés et non droite de régression. La raison est que, historiquement, on a défini la régression comme étant la courbe (pas nécessairement une droite) représentant E[Y|X]. Cette courbe n'est une droite, assurément, que lorsque les variables X et Y suivent conjointement une loi binormale. Dans les autres cas, la droite des moindres carrés est la meilleure approximation linéaire (meilleure au sens des moindres carrés) que l'on puisse faire de la courbe E[Y|X].

Une autre situation où la courbe est une droite se produit lorsque la variable X est un paramètre que l'on peut contrôler. Il suffit alors que les résidus du modèle suivent une loi normale de moyenne nulle pour que E[Y|X] coïncide avec une droite. En sciences de la terre, toutefois, il est relativement peu fréquent que l'on puisse vraiment contrôler des variables.

Remarque: Une régression peut être significative ou non selon la force du lien linéaire (corrélation) qui unit les deux variables. Le modèle adopté, même significatif, peut présenter un manque d'ajustement important (i.e. le modèle n'est pas le bon modèle).

Exemple numérique: L'exemple suivant est tiré de Krumbein and Graybill (1965), pp. 237-241. On cherche à établir la relation existant entre le degré d'arrondi (variable à expliquer Y) et la taille de galets de plage (variable explicative X).

# échantillon	degré d'arrondi (y)	Taille du galet (mm) (x)
1	.62	52
2	.74	43
3	.65	36
4	.71	32
5	.68	27
6	.59	26
7	.49	22
8	.67	37
9	.64	24
10	.56	19
11	.51	13

de ces données, on calcule les quantités suivantes:

b_0=.4903
b_1=.00443
Σe^2=.0382 Σe=0
$\Sigma(y-y_m)^2$=0.0063 y_m est la moyenne de y
$\Sigma(y_p-y_m)^2$=0.0025

Discussion: Bien que l'on puisse montrer que la régression est significative, ce modèle n'explique que 40% (.0025/.0063) de la variation de Y (arrondi). De plus ce modèle prédit des arrondis supérieurs à 1 pour X>115 mm, ce qui est physiquement impossible. Un modèle basé sur l'équation différentielle suivante serait peut-être préférable:

$$\frac{dR}{dX} = a(R_0 - R) \qquad 2.8$$

où R_0 est la limite d'arrondi possible (1 par exemple)
R est l'arrondi
X est la taille des galets.

Cette équation exprime que l'arrondi augmente à un taux décroissant en fonction de la taille des galets. En solutionnant cette équation différentielle et en imposant que pour X=0 on ait R=0, on trouve alors la relation suivante:

$$-\ln\left[\frac{R_0 - R}{R_0}\right] = aX \qquad 2.9$$

Il s'agit bien d'une équation linéaire que l'on estime par la méthode des moindres carrés. Toutefois, à la différence de tantôt, on doit imposer que la droite passe par l'origine. Le coefficient "a" est alors obtenu en solutionnant:

$$a = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \qquad 2.10$$

où y_i désigne $-\ln((R_0-R_i)/R_0)$

Une fois "a" obtenu, on estime R par:

$$R = R_0[1 - e^{-aX}] \qquad 2.11$$

Remarque: La droite obtenue est la droite des moindres carrés dans l'espace de la variable transformée Y. Ceci ne garantit pas que la courbe obtenue par transformation inverse dans l'espace de R soit la courbe des moindres carrés. Pour cette raison, autant que possible, on essaie de ne pas transformer la variable Y, mais plutôt les variables X. Ici, cela n'était pas possible.

Bien que le modèle soit plus acceptable physiquement, il fournit de moins bonnes estimations de l'arrondi. On obtient en effet les quantités suivantes pour les erreurs de prédiction:

Σe^2=.115 Σe=.004.

La somme des erreurs au carré est supérieure à celle observée pour le modèle linéaire. Le modèle semble aussi indiquer un léger biais (somme des erreurs différentes de 0). Ce biais est causé par la transformation requise pour obtenir un estimé de R. On conclut qu'il faut être prudent lorsqu'on effectue la régression linéaire sur une variable transformée, la transformation inverse pouvant causer plusieurs problèmes. Autant que possible, on évitera de transformer la variable Y. Si c'est nécessaire en raison de la nature des données, on vérifiera que la solution, après transformation inverse, conserve de bonnes propriétés (somme des carrés des erreurs, biais faible, etc.). Si nécessaire, des ajustements seront alors faits au modèle.

2.4 Régression linéaire multiple

Dans cette section, nous généralisons et étendons les résultats précédents au cas plus intéressant où l'on cherche à expliquer une variable Y par un ensemble de variables X. De façon à simplifier la notation, on utilisera la notation matricielle (voir annexe A).

Soit une variable Y que l'on veut relier à p variables X par le modèle linéaire suivant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p + \varepsilon \qquad 2.12$$

On cherche à estimer les p+1 coefficients $\beta_0, \beta_1, ..., \beta_p$ de façon à minimiser le carré de l'erreur "e" commise.

Plaçons nos "n" observations en colonne dans un vecteur et les n observations des X dans une matrice. L'équation précédente s'écrit alors:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & X_{1p} \\ 1 & X_{21} & X_{22} & . & . & X_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & X_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix} \qquad 2.13$$

Ou, plus simplement:

$$Y = Xb + e \qquad 2.14$$

La somme des carrés des erreurs s'écrit:

$$SCE = e'e = (Y - Xb)' (Y - Xb) \qquad 2.15$$

On voit que SCE est une fonction des "b". On les choisira de façon à minimiser SCE. Le minimum de SCE est atteint lorsque toutes les dérivées partielles de SCE par rapport aux différents b_i s'annulent:

$$SCE = Y'Y - Y'Xb - b'X'Y + b'X'Xb \qquad 2.16$$

$$\frac{\delta SCE}{\delta b} = 0 = (X'X)b - X'Y \qquad 2.17$$

d'où on tire finalement:

$$b = (X'X)^{-1} X'Y \qquad 2.18$$

Ce système de p+1 équations à p+1 inconnues est appelé **"équations normales"** de la régression.

Exercice 1: Si p=1, démontrez que le système d'équations normales permet de retrouver les résultats énoncés précédemment dans le cas de deux variables.

Question 5: Comment faudrait-il modifier la matrice X pour tenir compte du cas de la régression passant par l'origine?

Remarque: Lorsque p=1, la régression définit une droite. Lorsque p=2, un plan de régression est défini. Lorsque p=3, un hyperplan est défini, de même pour p>3.

2.4.1 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	$e'e$ $(Y-Y_p)'(Y-Y_p)$	Σe_i^2	n-(p+1)	$SCT=SCR+SCE$ $SCT_m=SCR_m+SCE$ $SCE \perp SCR$ $SCE \perp SCR_m$

Note: Y_p= Xb ; i.e. valeurs prédites par la régression.
Y_m= vecteur n x 1 ayant la moyenne de Y à chaque entrée.

Remarque: Dans ce tableau, d.l. signifie degrés de liberté. Pour comprendre d'où viennent ces degrés de liberté, il faut savoir que toutes les sommes de carrés précédentes peuvent se mettre sous la forme quadratique Y'AY où la matrice A est une **matrice idempotente** (rappel: une matrice idempotente est une matrice telle que A*A=A). Le rang de la matrice A définit le nombre de degrés de liberté associés à la forme quadratique. Les degrés de liberté correspondent donc à la dimension de l'espace associée à la somme des carrés (nombre d'éléments non linéairement dépendants dans la somme des carrés). Deux formes quadratiques (somme de carrés) sont orthogonales si les matrices idempotentes les définissant sont orthogonales.

Exemple: $SCE=e'e = (Y-Xb)'(Y-Xb)$
$= (Y-X(X'X)^{-1}X'Y)'(Y-X(X'X)^{-1}X'Y)$

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

$SCE=Y'(I-M)Y$; on vérifie que I, M et (I-M) sont des matrices idempotentes.

$SCR=Y_p'Y_p = (Xb)'(Xb)$
$= (MY)'(MY)$
$= Y'MY$

On a M(I-M)=0 : les deux sommes de carrés sont orthogonales.

Note: La matrice M est appelée "hat matrix" en anglais. Le nom vient du fait que l'on peut écrire : $\hat{Y}=Xb=X(X'X)^{-1}X'Y=MY$. Cette matrice apparaît dans plusieurs résultats concernant la régression (matrice de variances-covariances des résidus, somme des carrés, projections, etc.)

Exercice 2: Exprimez chacune des sommes de carrés du tableau précédent sous la forme Y'AY. Vérifiez que les matrices sont idempotentes et vérifiez les orthogonalités décrites. (note: pour certaines démonstrations, on utilisera le fait que M 11'/n = 11'/n)

Exercice 3: Démontrez les égalités suivantes:

$e'Y_m=0$
$e'Y_p=0$
$e'\mathbf{1}=0$ $\mathbf{1}$ est un vecteur de 1
$Y'Y_p=Y_p'Y_p$

2.4.2 Tests statistiques en régression

Les tests statistiques utilisés en régression reposent sur l'hypothèse d'une distribution normale des résidus, de même variance et moyenne, et indépendante. Étant donné que l'on a Y=Xb+e, que X est considéré comme un paramètre que l'on peut fixer, que b est un vecteur de constantes, il suit que la distribution de Y peut être déduite uniquement de la distribution des résidus. Également, on notera que les formes quadratiques Y'AY se résument en quelque sorte à des sommes pondérées de carrés de variables normalement distribuées. On ne sera pas surpris, dans ces circonstances de voir apparaître des lois du Khi-deux et de Fisher pour définir les tests en régression. A cet effet, deux théorèmes sont fondamentaux:

Théorème 1: Si $Y \sim N(u,\sigma^2 I)$ alors $Y'AY/\sigma^2 \sim \chi^2_{rangA,\delta}$ si et seulement si A est une matrice idempotente. (note: δ est un paramètre de non-centralité relié au fait que $E[Y] = \mu = Xb \neq 0$. δ vaut $(\mu'A\mu)/\sigma^2$; si $\mu=0 \rightarrow \delta=0$).

Théorème 2: Si $Y \sim N(u,\sigma^2 I)$ alors les formes quadratiques $Y'AY/\sigma^2$ et $Y'BY/\sigma^2$ où A et B sont des matrices idempotentes, sont distribuées indépendamment si et seulement si AB=0 (i.e. A est orthogonale à B).

Rappels: i. Une somme de carrés de n variables aléatoires indépendantes et distribuées suivant une N(0,1) est distribuée suivant une χ^2_n.

ii. Soit $Y \sim \chi^2_n$ et $Z \sim \chi^2_m$ et Y est indépendante de Z. Alors $(Y/n) / (Z/m) \sim F_{n,m}$. Le rapport de deux chi-deux indépendantes est distribué suivant une loi Fisher.

On a maintenant tous les éléments nous permettant de construire des tests statistiques. Il suffit de déterminer quelles sont les formes quadratiques parmi les différentes sommes des carrés qui répondent aux énoncés des théorèmes 1 et 2.

Rappel sur les tests statistiques :

Un test statistique consiste à confronter les résultats d'une expérience à une hypothèse de départ (H_0). Pour réaliser un test, il faut connaître la distribution d'une statistique en supposant l'hypothèse de départ vérifiée.

Nous nous concentrerons sur le test le plus important en régression: "Est-ce que la régression explique quelque chose (une fois enlevé l'effet de la moyenne)", i.e. est-ce que la pente de la régression est significativement différente de zéro (cette pente est égale à zéro lorsqu'il n'y a pas de relation entre les variables Y et X). Si les variables X expliquent vraiment Y alors SCE (somme des carrés des erreurs) sera faible car les erreurs seront faibles et SCR_m sera élevée. On cherchera donc à construire une statistique à partir de ces deux éléments, dont on connaîtra la distribution. Les théorèmes précédents seront ici utilisés.

Supposons que les erreurs "ε" du modèle suivent une distribution $N(0,\sigma^2 I)$. Ceci entraîne que $Y \sim N(X\beta,\sigma^2 I)$. On peut montrer que $SCR_m = Y'(M-11'/n)Y$. La matrice $(M-11'/n)$ est une matrice idempotente. Utilisant le théorème 1, il découle que SCR_m/σ^2 est distribué suivant une $\chi^2_{p,\delta}$ car $(M-11'/n)$ est une matrice idempotente de rang p.

De la même façon, on trouve que SCE/σ^2 est distribué suivant une $\chi^2_{(n-(p+1))}$ car $SCE=Y'(I-M)Y$ et $(I-M)$ est une matrice idempotente de rang $(n-(p+1))$. Ici, le paramètre de non-centralité $\delta=0$ car $\beta'X'(I-M)X\beta=0$ (Note: $X\beta=E[Y]$).

Soit $H_0 : \beta_1=\beta_2=.....=\beta_p=0$; i.e. la régression est nulle, toutes les pentes du modèle sont égales à zéro.
vs H_1 : non H_0 ; i.e. une pente au moins est différente de zéro; la régression explique quelque chose.

Sous H_0, SCR_m/σ^2 est distribué suivant une χ^2_p (i.e. le paramètre $\delta=0$). Utilisant le théorème 2, on trouve que $(SCR_m/p) / (SCE/(n-(p+1))$ est distribué suivant une loi $F_{p,(n-(p+1))}$ car on a $(I-M)(M-11'/n)=0$ et les deux lois χ^2 sont donc indépendantes.

On calcule le rapport précédent que l'on compare à la valeur F lue dans la table. Si le rapport est supérieur à la valeur critique de la table c'est que la régression explique quelque chose et par conséquent on doit rejeter H_0.

Exercice 4: Construisez le test pour vérifier si la moyenne de Y est significativement différente d'une valeur "m" donnée (m pouvant être 0).

Exercice 5: Construisez le test pour vérifier si la régression, globalement, explique quelque chose (incluant la moyenne).

Exemple numérique: On a effectué la régression de Y sur X_1 et X_2 avec 13 observations. On a obtenu $SCR_m = 30$ et $SCE = 50$. La régression est-elle significative?

On calcule $(30/2) / (50/10) = 3.0$
On lit $F_{2,10} = 4.10$ (au niveau $\alpha=0.05$)
On conclut que la régression n'est pas significative (au niveau $\alpha=.05$)

Le tableau suivant présente les principales propriétés de la régression en fonction du niveau d'hypothèses nécessaire pour les obtenir.

Hypothèse sur ε (modèle)				
Élement	Aucune	$E[\varepsilon]=0$	$E[\varepsilon]=0$ $Var[\varepsilon]=\sigma^2 I$	ε normal
b (estimé)	$b=(X'X)^{-1}X'Y$	$E[b]=\beta$ (modèle)	$Var[b]=\sigma^2(X'X)^{-1}$	b normal
Y_p	$Y_p=Xb$	$E[Y_p]=E[Y]=X\beta$	$Var[Y_p]=\sigma^2 M$	Y_p normal
Y	$Y'Y_p=Y_p'Y_p$	$E[Y]=X\beta$	$Var[Y]=\sigma^2 I$	Y normal
e (estimé)	$e=Y-Xb$ $1'e=0$ $X'e=\mathbf{0}$ $Y_p'e=0$	$E[e]=0$	$Var[e]=\sigma^2(I-M)$ $E[e'e]=\sigma^2(n-p-1)$	e normal

Note: $M=X(X'X)^{-1}X'$; dans le tableau, on estimera σ^2 par $s^2=SCE/(n-p-1)=CME$.

A l'aide de ce tableau, on peut construire les intervalles de confiance et les tests sur tout élément d'intérêt puisqu'on en connaît la distribution statistique.

Exemples:

i. Vous avez effectué la régression de Y en fonction de p variables X. Supposons que vous observez un nouvel ensemble de p valeurs soit $(1,x_1,x_2....x_p)=\mathbf{x_i}$. Vous calculez $Y_{pi}=\mathbf{x_i}b$. Construisez l'intervalle de confiance autour de Y_{pi} pour la valeur Y_i que vous devriez observer associée à ce $\mathbf{x}_i$.

On a $Var(Y_i-Y_{pi})=Var(Y_i)-2Cov(Y_i,Y_{pi})+Var(Y_{pi})$
La variance de Y_{pi} est $\sigma^2(\mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})$.
La variance de Y_i est σ^2.
La covariance entr Y_i et Y_{pi} est nulle puisque Y_{pi} est une combinaison linéaire des Y de la régression et que le Y_i n'a pas été utilisé dans la régression (les Y_i sont indépendants).
Donc $Var=\sigma^2(1+\mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})$.
Puisque σ^2 n'est pas connu, on le remplace par son estimateur SCE/(n-(p+1)) et on utilise une Student plutôt qu'une loi normale.

L'intervalle de confiance est donc:

$$Y_p \pm t_{n-(p+1),\alpha}\ s(1+\mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})^{0.5} \qquad \text{avec } s=[SCE/(n-(p+1))]^{0.5}$$

Remarque: À l'intérieur de la parenthèse, on reconnaît deux contributions différentes. Le premier terme représente la variation autour de la droite de régression, le second terme représente l'imprécision sur la position de cette droite de régression.

ii. Toujours dans le même contexte, vous fixez $\mathbf{x_i}$ et vous répétez plusieurs fois l'expérience (disons k fois). Quelle est l'intervalle de confiance pour la moyenne de ces k mesures?

Par un développement similaire, on arrive à:

$$Y_p \pm t_{n-(p+1),\alpha}\ s(1/k + \mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})^{0.5}$$

Remarque: Comme précédemment, on reconnaît deux termes différents, un tenant compte de la variance de la moyenne de k observations autour de la droite de régression, un tenant compte de l'incertitude sur cette droite de régression.

iii. L'intervalle de confiance pour la moyenne de Y, pour un vecteur x_i donné (i.e. $E[Y|x_i]$, ou ce qui est équivalent, la droite de régression), par:

$$Y_p \pm t_{n-(p+1),\alpha}\ s(\mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})^{0.5}$$

Remarque: Ici seul subsiste le terme d'incertitude sur la position de la droite de régression.

iv. Vous voulez construire un intervalle de confiance pour un coefficient ou un intervalle de confiance simultané pour un ensemble de coefficients.

1 seul coefficient: $b_i \pm t_{n-(p+1),\alpha}\ s_{bi}$ où s_{bi} est l'écart-type du coefficient b_i obtenu en prenant la racine carrée de $(X'X)^{-1}s^2$ à la position correspondante pour le coefficient sur la diagonale.

plusieurs coefficients: $(\beta\text{-}b)'X'X(\beta\text{-}b) \leq (p+1)s^2F_{p+1,n-(p+1),1-\alpha}$ où F est la loi de Fisher. Cette équation définit un ellipsoïde de confiance de niveau $1\text{-}\alpha$.

v. Vous effectuez deux régressions avec deux ensembles de données différents (mais avec les mêmes variables i.e. le même modèle) et vous voulez tester si les deux régressions peuvent être considérées comme étant identiques.

Voir section 2.4.7

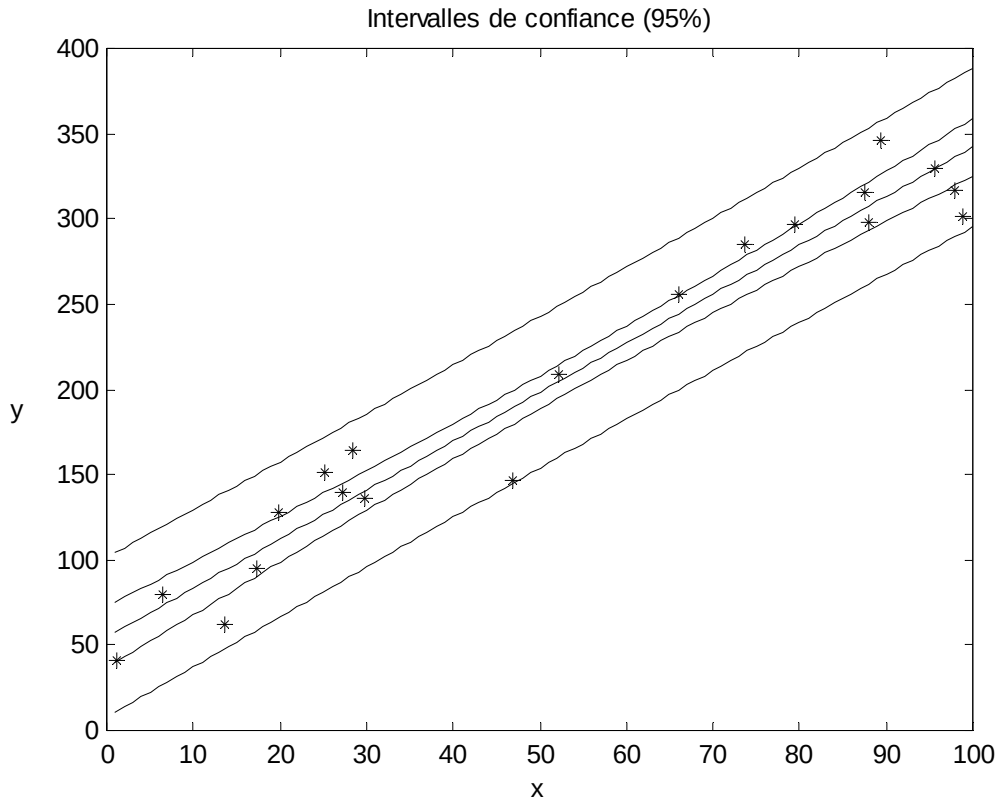
vi. Vous voulez vérifier si deux ou plusieurs coefficients de la régression sont égaux.

Voir section 2.4.7

vii. Vous voulez vérifier si une régression donnée suit un modèle spécifié.

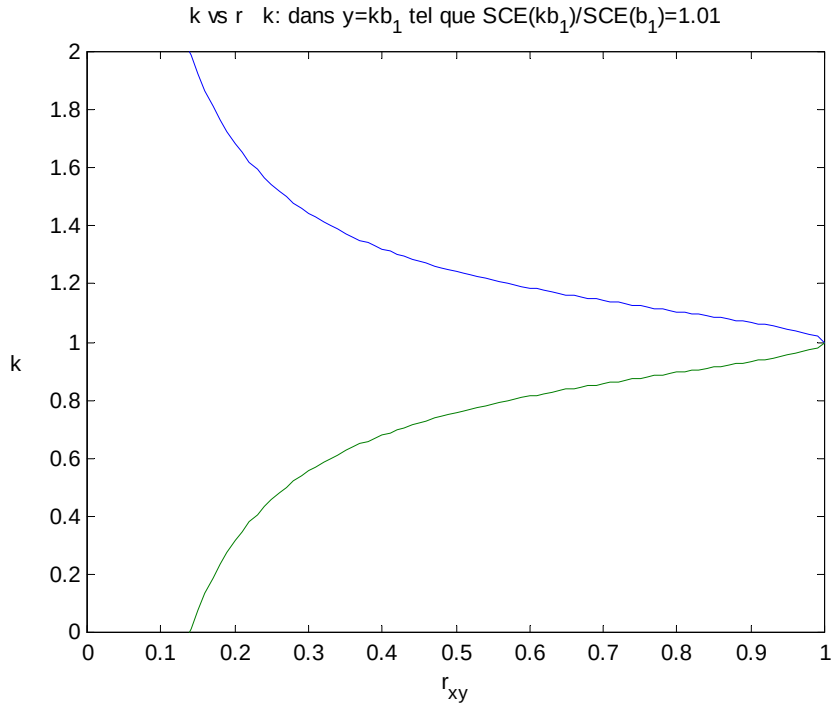
Voir section 2.4.7

La figure suivante montre un exemple de l'intervalle de confiance obtenu pour E[Y|X] (intervalle le moins large, cas iii. ci-dessus) et pour une observation de Y à X fixé (intervalle le plus large, cas i. ci-dessus). Notez comme l'intervalle de confiance est plus étroit près de la moyenne des X et plus large lorsqu'on s'éloigne de celle-ci. Ceci est dû à l'incertitude sur la pente réelle de la droite de régression.



2.4.2.1 Remarque sur le test de signification

Le fait que la régression soit significative ne veut pas dire qu'il s'agit du seul modèle acceptable, loin s'en faut. En fait plusieurs droites voisines de la droite de régression pourraient donner un ajustement presque aussi bon. Ainsi, si l'on compare un modèle $Y=b_0+b_1X_1+e$ au modèle $Y=b_0+kb_1X_1+e$, on peut exprimer le ration $SCE(kb_1)/SCE(b_1)$ en fonction de k et du coefficient de corrélation simple r_{xy}. Si l'on fixe le ratio à disons 1.01, l'on peut alors exprimer k en fonction de r_{xy}. C'est ce que montre la figure suivante. Comme on le voit, pour de faibles r_{xy}, il y a des droites fort différentes qui donneraient un ajustement quasi-équivalent.



2.4.3 Le coefficient de corrélation multiple (ou coefficient de détermination)

Le coefficient de corrélation multiple, noté R^2 représente la proportion de la variance totale de Y qui peut être prise en compte par les variables X. Lorsque le modèle de régression comporte une constante, on le définit comme:

$$R^2 = \frac{SCR_m}{SCT_m} \qquad 2.19$$

Lorsqu'il n'y a pas de constante dans le modèle où lorsqu'on veut pouvoir comparer 2 modèles dont l'un est élaboré directement sur Y et l'autre sur une transformation de Y, f (Y), on calcule la statistique suivante:

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2} = 1 - \frac{SCE}{SCT_m} \qquad 2.19b$$

Note: Plusieurs logiciels de régression donnent R^2=SCR/SCT pour le modèle sans constante. Ceci devrait être évité car il est alors impossible de comparer la performance du modèle avec constante et celle du modèle sans constante.

Note: Les deux expressions précédentes sont équivalentes dans le cas d'un modèle avec constante.

Note importante concernant les transformations : Lorsqu'on effectue une transformation sur Y (ex. log), et que l'on effectue la régression sur la variable transformée, le R^2 que donne les programmes est le R^2 pour la prédiction de la variable transformée. On ne peut donc pas le comparer avec un autre R^2 qui serait obtenu directement sur Y. Pour pouvoir comparer le pouvoir explicatif des deux modèles, il faut d'abord effectuer la transformation inverse sur Y et calculer le R^2 avec la relation 2.19b.

Question 6: Vous servant de la définition de R^2 et des résultats précédents concernant les tests, construisez un test pour déterminer le caractère significatif de R^2.

Question 7: Lorsqu'on a une seule variable explicative dans la régression, quel est le lien existant entre R^2 et le coefficient de corrélation simple r. Déduisez un test pour le coefficient de corrélation simple.

Remarque: Les tests statistiques sont valides uniquement si les postulats du modèle sont satisfaits, i.e. les résidus du modèle sont indépendamment et identiquement normalement distribués. Cependant, le fait qu'une régression soit significative ne dit pas grand chose sur la valeur du modèle trouvé. Tout ce que cela indique c'est que la relation observée ne peut être raisonnablement considérée comme le fruit du hasard. Pour que la relation établie soit de quelque utilité (pour des prédictions entre autres), il faut que R^2 soit considérablement supérieur au R^2 critique nécessaire pour obtenir un test positif. Certains auteurs recommandent un R^2 quatre fois supérieur au R^2 critique.

2.4.3.1 Quelques résultats spécifiques au cas de 2 variables

Item	*Formule générale*	*Cas avec p=1*
Coefficients de la régression :	$b=(X'X)^{-1}X'Y$	$b_0 = \bar{Y} - b_1\bar{X}$ $b_1 = \frac{s_{xy}}{s_x^2}$
Coefficient de corrélation multiple :	$R^2 = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = 1 - \frac{SCE}{SCT_m}$	$R^2 = r_{xy}^2$
Variances-covariances des coefficients :	$\sigma^2(X'X)^{-1}$	$\mathrm{Var}(b_0) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i-\bar{x})^2}$ $\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$ $\mathrm{Cov}(b_0,b_1) = -\frac{\bar{x}\ \sigma^2}{\sum(x_i-\bar{x})^2}$
Intervalles de confiance pour $E[Y\|x=x_0]$	$Y_p \pm t_{n-(p+1),\alpha}\ s(\mathbf{x_0}(X'X)^{-1}\mathbf{x_0}')^{0.5}$ Note : $s=CME^{1/2}$; t : Student	$Y_p \pm t_{n-2,\alpha}\ s\left\{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$

Note : σ^2 est estimé en pratique par CME=SCE/(n-p-1)

2.4.4 Validation du modèle de régression; étude des résidus

L'étude des résidus d'un modèle de régression vise plusieurs objectifs:

i. Vérifier les postulats du modèle: normalité, homogénéité des variances des résidus (homoscédasticité) et indépendance des résidus.

ii. Détecter des données aberrantes qui s'écartent considérablement du modèle.

iii. Détecter des tendances particulières (ex. comportement quadratique des résidus) et des relations des résidus avec des variables externes qui permettraient d'affiner le modèle.

La **normalité** se vérifie essentiellement en construisant l'histogramme ou la fréquence cumulée des résidus. On peut vérifier l'ajustement à une normale visuellement ou effectuer des test de normalité (ex. test d'ajustement du χ^2, test de Kolmogorov-Smirnov, etc...).

L'**indépendance des résidus** peut être testée en ordonnant les résidus en fonction d'un critère donné et en effectuant un test du genre: test des signes des résidus ou test de la corrélation entre résidus successifs dans la séquence ordonnée. Le test des signes (Draper et Smith, 1966; p.95) est un test non-paramétrique qui examine si l'arrangement des signes des résidus dans la séquence est aléatoire ou anormalement groupé ou encore anormalement fluctuant. Le test de corrélation consiste à calculer la corrélation entre les résidus et eux-mêmes décalés d'un pas dans la séquence. Si la corrélation est significative, alors il n'y a pas indépendance des résidus.

Le critère servant à ordonner la séquence peut être une variable interne (ex. une des variables X, la variable Y_p) ou une variable externe (ex. temps, collectionneur, laboratoire, provenance des échantillons, etc...)

Question 8: Suggérez un outil géostatistique (c.f. GLQ3401) qui permettrait d'évaluer visuellement l'indépendance des résidus.

L'**homogénéité des variances** des résidus se vérifie en ordonnant les résidus selon un critère comme ci-dessus et en vérifiant que les résidus montrent des variations de même amplitude pour toute la séquence ordonnée. Si ce n'est pas le cas, alors on peut tenter de corriger la situation à l'aide de transformations telles le logarithme ou la racine carrée qui ont habituellement pour effet de stabiliser la variance.

La **détection de données aberrantes** s'effectue en considérant les résidus qui s'écartent beaucoup de zéro. Les résidus situés à plus de trois écarts-types (note l'ecart-type des résidus est estimé par $(SCE/(n-p-1))^{0.5}$), sont suspects et doivent être examinés avec attention. Si des erreurs sont responsables de ces valeurs élevées, on doit les éliminer et reprendre la régression. Si aucune cause d'erreur ne peut les expliquer, alors il faut soit chercher à affiner le modèle pour mieux expliquer ces données, soit chercher de nouvelles observations avec les mêmes valeurs de X que ces données pour en vérifier la validité.

La **détection de tendances particulières** dans les données se fait en reportant sur des diagrammes binaires les résidus en fonction de chacune des variables X. Des diagrammes binaires entre les résidus et des variables externes peuvent suggérer l'inclusion de nouvelles variables ou la transformation de variables existantes dans le modèle afin d'en améliorer la performance.

Note: Comme les résidus ont théoriquement comme variance $\mathrm{Var}(e)=\sigma^2(I-M)$, il découle que la variance des résidus dépend des valeurs X correspondantes. Certains logiciels normalisent les résidus bruts en les divisant par l'écart-type ("studentised residuals").

2.4.4.1 Transformations des données

Transformation sur Y

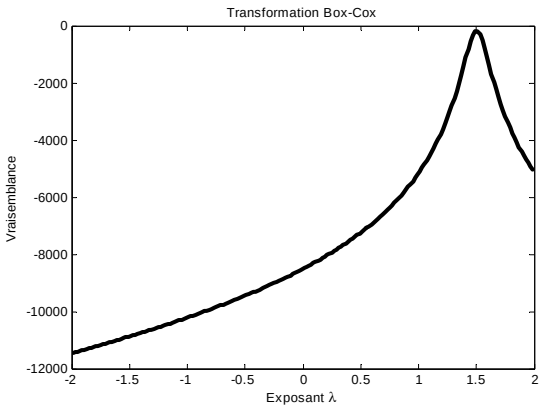
i. Si l'on doit transformer Y, normalement, il est préférable d'interpréter les résultats en terme de la variable transformée plutôt que de chercher à effectuer la transformation inverse. En effet, après transformation inverse, la régression n'est plus une droite, les erreurs ne sont plus symétriquement distribués autour de la "droite" de régression (i..e les erreurs suivent une distribution autre que normale), et la valeur transformée correspond à une médiane de la distribution, non à une espérance. De plus la somme des carrés des résidus sera supérieure à ce qu'il serait possible d'obtenir si l'on effectuait directement la régression non-linéaire sur Y.
ii. La transformation de Y devrait viser à accroître la normalité des résidus. Pour ce faire, la transformation de Box-Cox est fréquemment utilisée: $Z=Y^a$ avec « a » entre -2 et 2. Pour a=0, prendre Z=ln(Y). On doit estimer le paramètre "a". Pour ce faire, on calcule la régression de Z sur X pour un "a" donné. On obtient SCE(a) et on calcule la fonction de vraisemblance:

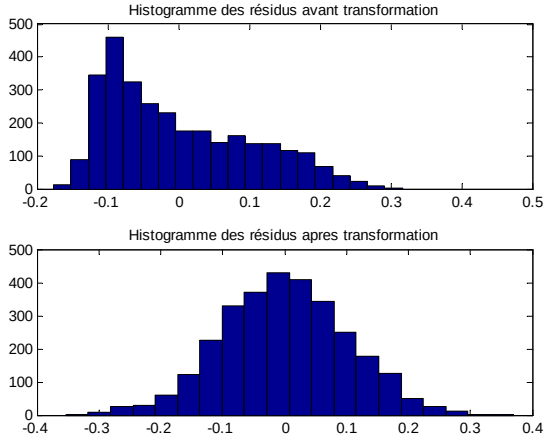
$$L(a)=n*\ln(|a|)-n/2*\ln(SCE(a))+n*(a-1)*\text{moyenne}(\ln(Y)) \quad \text{si } a\neq 0$$
$$L(a)=-n/2\ln(SCE(a)-n*\text{ moyenne}(\ln(Y)) \quad \text{si } a=0$$

On retient le "a" qui maximise la vraisemblance.

Exemple :

On simule 3000 ensembles de valeurs du modèle $Y^{1.5}=(1+2*X+e)$ avec e~N(0,0.01) et X uniforme sur [10 ,30].





On applique l'algorithme précédent et on trouve la valeur maximale à a=1.51, ce qui est très près de la vraie valeur 1.5. Les résidus de la régression Y=Xb+e montent une distribution qui n'est pas normale. L'inférence à propos de ce modèle est donc douteuse. Par contre les résidus du modèle $Y^{1.51}$=Xb+e sont distribués, visuellement, comme une normale.

Se rappeler toutefois, que si l'on adopte la transformation de Box-Cox, alors c'est la variable transformée qui est estimée. La transformation inverse fournit généralement un modèle non-optimal et même biaisé. Ainsi dans l'exemple précédent, la somme des carrés des erreurs vaut 31 pour la régression directe sur Y et vaut 127 pour la régression sur $Y^{1.5}$ puis transformation inverse.

Transformation d'un (ou plusieurs) X

Il faut considérer 2 cas:

i. Le maximum de Y est observé à l'intérieur du domaine de définition de X_i. Les transformations polynômiales de X_i peuvent alors être appropriées (i.e. inclure des termes en X_i^2, X_i^3....

ii. Le maximum de Y apparaît à une des extrémités d'un X_i donné. On peut envisager une transformation de puissance de type Box-Cox sur X_i. Weisberg (1985, p.153) décrit une méthode approximative pour déterminer si une telle transformation est requise et évaluer la puissance à utiliser.

2.4.4.2 La notion d'influence d'une observation

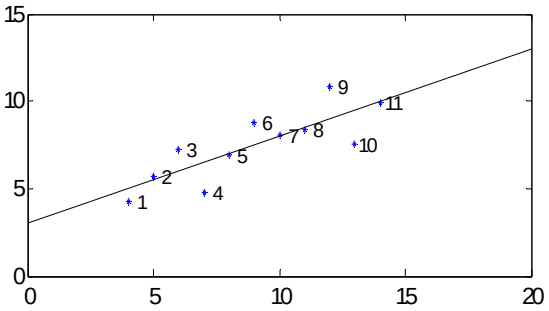
Lorsqu'on effectue une régression, il est important de vérifier si le modèle obtenu peut être causé par une (ou quelques unes) observation particulière. On espère habituellement que le modèle représente une caractéristique générale des données et non l'influence d'une seule donnée particulière. L'examen des résidus permet souvent d'identifier de telles données, mais ce n'est pas toujours le cas. L'idée générale est ici d'enlever de la régression chacune des observations à tour de rôle et d'examiner comment fluctuent les coefficients de la régression. Si le fait d'enlever une valeur change considérablement les coefficients de la régression, alors le modèle obtenu avec toutes les observations est fortement influencé par cette observation et il y a lieu de s'interroger sur sa validité.

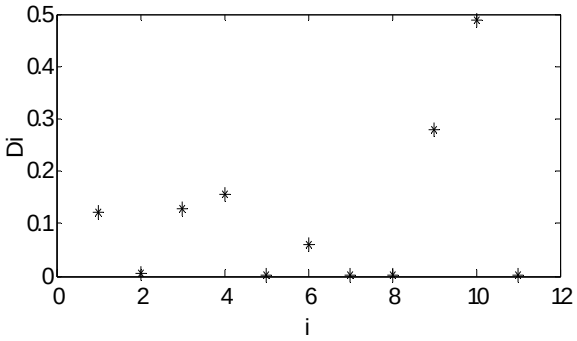
La figure suivante montre 4 ensembles de données ayant les mêmes coefficients "b", les mêmes R^2 et les mêmes CME. Pourtant, seul le 1er modèle est adéquat. On peut mesurer l'influence d'une observation à l'aide de la distance suivante (distance de Cook):

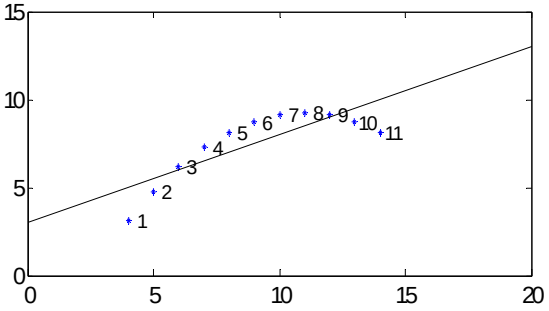
$$D_i = \frac{(b_{(i)} - b)'(X'X)(b_{(i)} - b)}{(p+1)CME} = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{(p+1)CME} \qquad 2.20$$

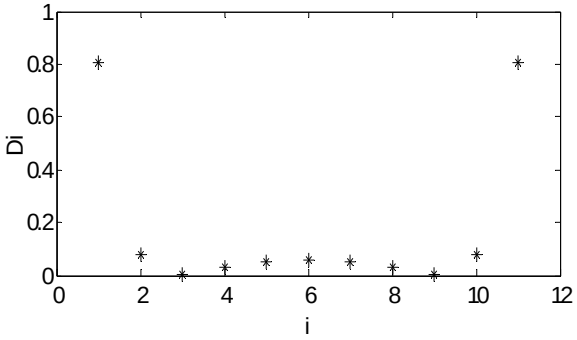
La notation (i) signifie que la $i^{ème}$ observation est enlevée. Weisberg (1985) indique que les observations présentant un D_i supérieur à 1 sont très influentes et doivent être examinées avec attention.

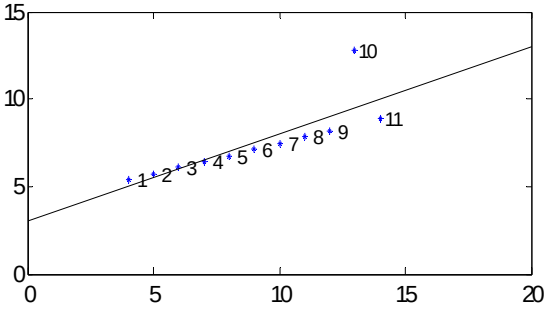
La figure montre, à droite, les D_i associées à chaque observation des données de gauche. Les 4 régressions ont exactement les mêmes coefficients b et le même R^2.

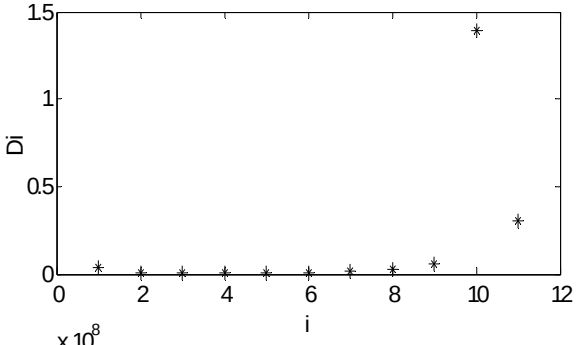

1
2
3
4
5
6
7
8
9
10
11

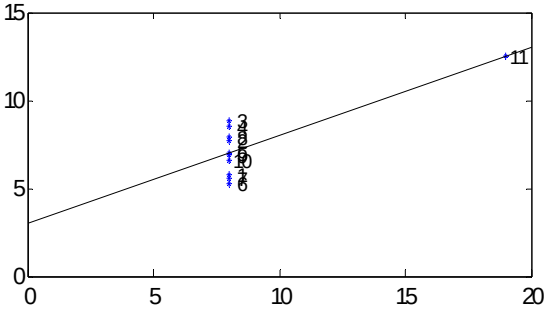

Di
i

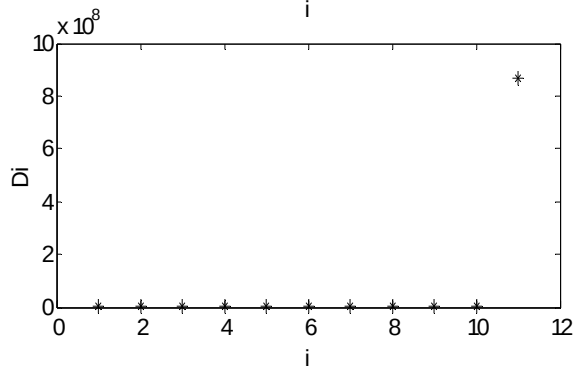

1
2
3
4
5
6
7
8
9
10
11


Di
i


1
2
3
4
5
6
7
8
9
10
11


Di
i


11


x 10^8
Di
i

2.4.5 Ajout d'une ou de plusieurs variables (complément sur les tests)

On peut ajouter des variables dans le modèle de régression en grande quantité. Il faut donc se donner un outil pour déterminer si l'ajout d'une ou de plusieurs variables améliore vraiment le modèle de régression.

Question 9: Quel est le nombre maximum de variables que l'on peut inclure dans une régression? Que se passe-t-il lorsqu'on atteint ce nombre? Que vaut alors le R^2?

Question 10: Soit un modèle donné auquel on ajoute une variable. Quelle relation pouvez-vous établir entre le nouveau R^2 et l'ancien?

Soit un modèle réduit (celui ayant le moins de variables):

$$Y = X_r \beta_r + \varepsilon_r \qquad 2.21$$

et un modèle complet constitué des mêmes variables que le modèle précédent auquel on ajoute "k" variables:

$$Y = X_c \beta_c + \varepsilon_c \qquad 2.22$$

Soit les sommes des carrés des erreurs des deux modèles.

$SCE_r = Y'(I-M_r)Y$
$SCE_c = Y'(I-M_c)Y$

La différence entre ces deux sommes de carrés s'écrit:

$SCE_r-SCE_c = Y'(M_c-M_r)Y$

On peut montrer que $M_cM_r = M_r$.

De ceci découle deux faits importants:

i. (M_c-M_r) est une matrice idempotente: donc la différence entre les sommes des carrés des erreurs suit une loi du χ^2 dont le nombre de degrés de liberté est donné par le rang de cette matrice qui est égal au nombre de variables ajoutées (k).

ii. $(M_c-M_r)(I-M_c) = 0$: donc la différence entre les sommes des carrés des erreurs est orthogonale à la somme du carré des erreurs du modèle complet. On sait que la somme des carrés des erreurs suit une loi du χ^2 dont le nombre de degrés de liberté est n-p-1, ou p est le nombre de variables dans le modèle complet.

Par conséquent (c.f. théorème 2):

$$\frac{(SCE_r - SCE_c)/k)}{SCE_c/(n-p-1)} \sim F_{k,(n-p-1)} \qquad 2.23$$

où p: nombre de variables dans le modèle complet.
k: nombre de variables ajoutées au modèle réduit.

Exemple numérique:

On a 13 observations pour lesquelles la régression de Y sur X_1 et X_2 a donné: SCE = 57.9

En ajoutant X_3 à la régression, on a obtenu SCE = 48.0.
Valait-il la peine d'ajouter X_3?

On calcule : [(57.9-48.0)/1] / [48.0/(13-4)] = 1.86
Dans une table, on lit $F_{1,9}$ = 3.36 (au niveau α=.10). On considèrera donc que X_3 n'ajoute rien à la régression une fois X_1 et X_2 inclus.

Note: Quand le test d'ajout porte sur une seule variable, le test F précédent est rigoureusement équivalent au test de Student sur le coefficient pour vérifier s'il est significativement différent de zéro.

2.4.5.1 Sélection optimale de variables

Souvent on a à notre disposition un nombre considérable de variables. On est alors intéressé à sélectionner parmi ces variables un sous-ensemble optimal de variables qui expliqueront presqu'autant la variable Y que l'ensemble complet des variables. Différentes techniques sont disponibles: sélection avant, élimination arrière, "stepwise". Une autre technique consiste a examiner les résultats de tous les sous-ensembles possibles de variable. Cette technique est évidemment prohibitive pour "p" trop grand.

Question 11: Dans un ensemble de p variables, combien y a-t-il de sous ensembles possibles?

i. Sélection avant: on démarre avec aucune variable dans la régression; à chaque itération, on introduit dans la régression la variable apportant la plus forte croissance du R^2. On arrête lorsque l'ajout de la variable n'amène plus d'augmentation significative du R^2 (ou de diminution de SCE).

ii. Élimination arrière: on démarre avec toutes les variables dans la régression; à chaque itération, on enlève la variable donnant la plus faible diminution du R^2. On arrête lorsque la diminution du R^2 (ou l'augmentation de SCE) devient significative.

iii. "Stepwise": on applique en alternance une itération de sélection avant et une itération d'élimination arrière. On arrête lorsqu'on ne peut ajouter une variable ni en éliminer une.

Note: Les résultats d'une sélection de variables ont tendance à surestimer fortement la qualité d'une régression. En effet, supposons Y et "p" variables X indépendantes entre elles et indépendantes de Y. Supposons un niveau α=.05 utiliser pour choisir une variable dans la régression. La probabilité qu' aucune variable n'entre dans la régression par une procédure de sélection est $(1-.05)^p$. Si p=30, cette probabilité n'est que de 0.21. Si p=50 elle devient .08, i.e. que l'on est alors presque certain de trouver une variable passant le test de signification même si en réalité aucun lien n'existe. Une règle simple est de choisir $\alpha'=\alpha/p$ comme niveau de choix d'une variable pour obtenir un niveau global de α. Ainsi $(1-.05/50)^{50}\approx 0.95$, $(1-.05/30)^{30}\approx 0.95$.

2.4.5.2 Sélection d'un sous-ensemble optimal de variables; la statistique C_p

La statistique C_p aide à choisir entre différents modèles candidats un modèle qui compte peu de paramètres tout en fournissant une bonne explication de Y. Cette statistique est assez directement reliée au R^2 sauf qu'une pénalité est inclue pour tenir compte du nombre de paramètres dans le modèle.

Soit un modèle complet ayant "c" variables et donc c+1 paramètres, et un modèle réduit ayant "p" variables et donc p+1 paramètres, on peut définir C_p des deux façons **équivalentes** suivantes (bien d'autres expressions pourraient également être dérivées, en fonction de R^2 par exemple):

$$C_p = \frac{SCE_p}{CME_c} + 2(p+1) - n \qquad 2.24$$

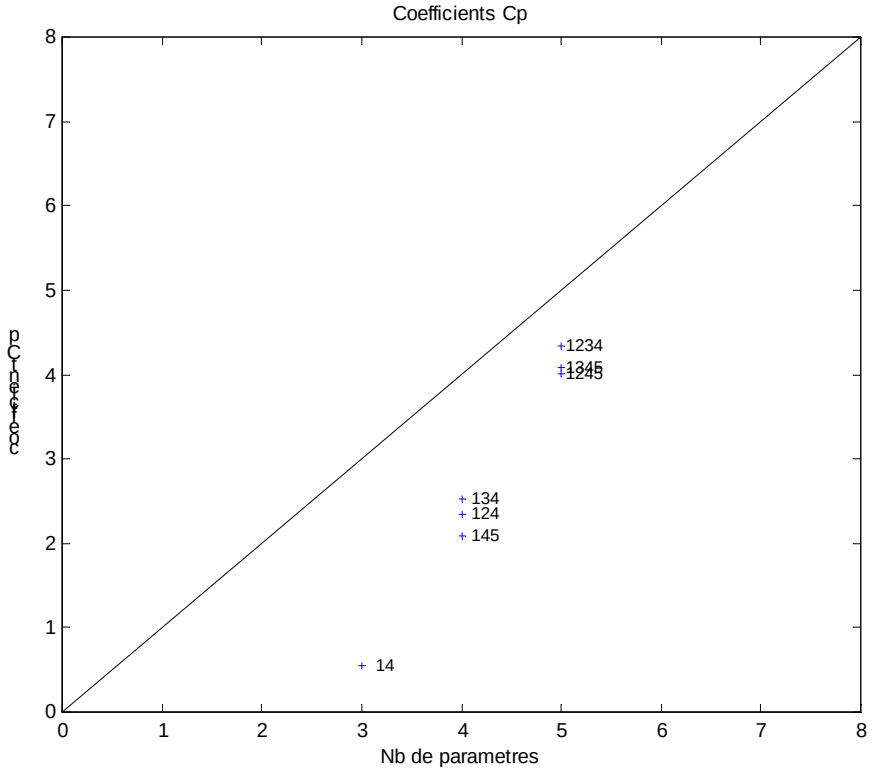
$$C_p = \frac{SCE_p - SCE_c}{CME_c} + (p+1) - (c-p)$$

où n est le nombre d'observations, "p" est le nombre de variables dans le sous-ensemble testé, "c" est le nombre de variables du modèle complet (p<c). Plus C_p est faible, meilleur est le modèle. Tous les modèles ayant C_p<p+1 (i.e. C_p<nb. de paramètres) sont potentiellement de bons modèles. À R^2 presqu'égaux, C_p favorise le modèle ayant le moins de variables.

La figure suivante montre le résultat d'une simulation. On a généré 5 variables aléatoires (X_1 à X_5) et on a construit $Y = X_1 + 2\,X_4 + e$. On a calculé les C_p pour les (2^5-1) sous-ensembles possibles de variables aléatoires et on les a représentés (la plupart des sous-ensembles montraient des valeurs de C_p de l'ordre de centaines ou de milliers). Seuls les sous-ensembles incluant les variables X_1 et X_4 ont montré de faibles C_p.

Le plus faible C_p est obtenu pour le sous ensemble n'incluant que X_1 et X_4. Ceci est toutefois un résultat particulier de cet ensemble de valeurs, puisqu'en effectuant d'autres simulations, il arrivait assez fréquemment que le couple X_1 X_4 ne présentait pas le plus faible C_p. Presque toujours cependant, le C_p pour X_1 et X_4 était parmi les plus faibles.

Remarque: Quand le nombre de variables est trop grand, il n'est pas possible de considérer tous les sous-ensembles de variables. Plusieurs stratégies ont été développées afin de n'avoir à considérer qu'un nombre restreint de sous-ensembles tout en effectuant une bonne sélection des variables.



2.4.5.3 "Trend Surface Analysis"

Une des premières techniques de cartographie à avoir été développée est le "trend surface analysis" qui n'est rien d'autre qu'une régression. Supposons que vous ayez prélevé un certain nombre de roches volcaniques (en Abitibi par exemple) pour lesquelles vous analysez le contenu en Na_2O. Vous pourriez être intéressé à produire une carte illustrant les grandes tendances dans la variation spatiale du Na_2O (note: on sait que des halos de lessivage en Na_2O accompagnent généralement la formation des gisements volcanogènes de cuivre). Cette carte illustrerait en quelque sorte le niveau de fond du Na_2O et les résidus ($Y-Y_p$) négatifs seraient alors marqueurs de zones potentiellement d'intérêt. La carte des valeurs prédites, quant à elle, devrait suivre, au moins grossièrement, la géologie connue.

Une telle carte peut être obtenue par régression à l'aide du modèle suivant où x et y représentent cette fois des coordonnées géographiques:

$$\%Na_2O = \beta_{00} + \beta_{10}x + \beta_{01}y + \beta_{20}x^2 + \beta_{11}xy + \beta_{02}y^2 + \ldots + \beta_{0k}y^k + \varepsilon \qquad 2.25$$

Cette équation exprime que le $\%Na_2O$ est vu comme un polynôme d'ordre k des coordonnées x et y. La détermination du degré optimal du polynôme peut être faite par la technique présentée précédemment, i.e. on augmente le degré du polynôme jusqu'à ce que le passage à un degré supérieur n'ajoute qu'une contribution non-significative.

Remarque: Très à la mode au début des années 60, cette technique n'est plus utilisée en cartographie; le krigeage, entre autres, lui étant supérieur. Son utilisation peut être encore valide pour filtrer d'un signal des éléments régionaux. En géophysique par exemple, on peut s'en servir comme étape préliminaire au traitement fréquentiel pour éliminer une dérive. On peut aussi l'utiliser dans cette optique, en géostatistique, préalablement au calcul de variogramme et au krigeage.

Question 12: Supposons que vous adoptiez un polynôme d'ordre élevé pour la cartographie du Na_2O dans l'exemple précédent. A quel danger sommes-nous exposés lorsque nous estimons la valeur de Na_2O à une bonne distance des points connus (nos observations)?

Exemple numérique:

L'exemple suivant montre un trend surface du Na_2O contenu dans 126 roches volcaniques prélevées dans la région de Normétal. On peut construire le tableau suivant pour déterminer le degré du polynôme que l'on doit retenir.

Degré	SCR_m d.l.	SCE d.l.	F (ajout)	F (table) α=.05	Décision
1	45.2 2	258.7 123	10.75	3.07	Signif.
2	82.0 5	221.8 120	6.64	2.68	Signif.
3	112.8 9	191.1 116	4.67	2.45	Signif.
4	133.9 14	170.0 111	2.76	2.30	Signif.
5	163.1 20	140.7 105	3.63	2.20	Signif.
6	178.3 27	125.6 98	1.69	2.13	Non-signif

Ici, on retiendrait donc un polynôme d'ordre 5.

Remarques:

i. Les cartes obtenues permettent de dégager les grandes tendances régionales. Elles sont habituellement de piètre qualité en ce qui concerne les phénomènes locaux. Pour ceux-ci, on préfèrera des méthodes mieux adaptées tel le krigeage utilisé en géostatistique.

ii. Les valeurs prédites aux points expérimentaux ne coïcident pas avec les valeurs observées. Plusieurs méthodes permettent de retrouver lors des prédictions les valeurs observées aux points expérimentaux (dont le krigeage). On dit alors que la régression n'est pas un interpolateur exact (le krigeage l'est).

iii. On ne doit jamais extrapoler en dehors de la zone couverte par les observations. Un polynôme d'ordre élevé definit une surface de prédiction qui a toutes les chances de diverger dès que l'on quitte le champ couvert par les données.

iv. On se méfiera des polynômes d'ordre élevé. Le nombre de paramètres à estimer augmente rapidement et surtout on se retrouve avec des variables dont l'ordre de grandeur est très différent. Tout ceci cause d'énormes problèmes de précision numérique et de stabilité des résultats.

v. On choisira habituellement de retenir le polynôme d'ordre k tel que le test s'avère négatif pour les ordres k+1 et k+2. Cependant pour k assez grand (5 ou 6) on arrêtera dès le premier test négatif.

vi. On doit se rappeler que pour que les tests soient valables il faut que les erreurs ε soient indépendantes les unes des autres. Ceci devrait être vérifié. Il y a fort a parier que très souvent cette hypothèse d'indépendance des résidus ne tient pas (surtout si l'on n'a pas le bon degré de polynôme). En effet, dans ce cas les résidus se regrouperont sur une carte selon un arrangement clairement non aléatoire. Lorsqu'on a le bon degré de polynôme, les résidus devraient présenter un caractère très erratique lorsque portés sur une carte et ceci peut nous guider pour choisir le degré du polynôme. On conservera à l'esprit que puisque les résidus (règle générale) ne sont pas indépendants, les tests seuls ne peuvent suffire à déterminer le degré du polynôme.

2.4.5.4 Application: correction géométrique de photos aériennes

Les photos aériennes et images de satellites (télédétection) souffrent très souvent de distorsions de l'image dues à des mouvements de la plate-forme, des perturbations atmosphériques, des défauts du capteurs et d'autres causes. Si on veut pouvoir superposer ces images sur un modèle de terrain (S.I.G. : système d'information géographique), il faut, au préalable corriger ces distorsions. Une des techniques possible pour ce faire est la régression; elle consiste à:

i. Identifier sur une carte de base, exempte de distorsions, une série de points de contrôles facilement repérables sur l'image à corriger. Noter les coordonnées (u_i, v_i) de ces points sur la carte de base et les coordonnées (x_i, y_i) sur la carte à corriger.

ii. Les coordonnées sur la carte de base sont les variables X de la régression. Les coordonnées de l'image à corriger sont les variables Y de la régression.

iii. On effectue deux régressions séparées (une pour chaque coordonnée de l'image à corriger). Le modèle de prédiction est un polynôme construit avec les coordonnées (u_i, v_i) de la carte de base qui fournit une valeur (x_i^*, y_i^*).

iv. En tout point (u_0, v_0) de la carte de base, on calcule avec l'équation de prédiction le point (x_0^*, y_0^*). La valeur sur l'image est lue et est représentée aux coordonnées (u,v). On obtient ainsi notre image corrigée.

2.4.6 Utilisation de variables indicatrices ("dummy variables")

Souvent, en plus de l'information purement quantitative à partir de laquelle on veut construire notre régression, on a à notre disposition une foule d'informations qualitatives que l'on voudrait bien incorporer dans notre modèle afin de le bonifier. Cette information qualitative pourrait être, à titre d'exemple:

- types de roches différents.
- textures différentes.
- mois, saison, année de prélèvement.
- techniques d'analyse, échantillonneurs, laboratoires différents.
- présence d'une faille séparant nos observations en deux groupes.
- machinerie, procédés utilisés.
- etc.

Le contexte, la connaissance que l'on a du phénomène, l'expérience et le jugement permettront à l'ingénieur d'identifier les facteurs qualitatifs pouvant influencer le modèle. L'étude minutieuse des résidus peut indiquer des lacunes du modèle et suggérer l'inclusion de variables qualitatives pour l'améliorer.

Ces variables qualitatives peuvent altérer le niveau de Y, la variabilité de Y, la droite de régression. Elles peuvent agir isolément ou se combiner à d'autres variables qualitatives ou quantitatives.

Exemple: Soit une régression à deux variables. Supposons que l'on a deux types de roches différents. On code une variable indicatrice:

I=0 si roche de type 1
I=1 si roche de type 2

Pour permettre une ordonnée à l'origine différente dans le modèle, en fonction du type de roche, on écrit:

$$Y = b_0 + b_1 I + b_2 X + e$$

Pour le type 1, l'ordonnée sera: b_0
Pour le type 2, l'ordonnée sera: $b_0 + b_1$

Pour permettre une ordonnée commune mais une pente différente, on écrira:

$$Y = b_0 + b_1 I X + b_2 X + e$$

Pour le type 1, la pente sera: b_2
Pour le type 2, la pente sera: $b_1 + b_2$

Pour permettre deux droites de régression différentes selon le type de roche:

$$Y = b_0 + b_1 I + b_2 I X + b_3 X + e$$

Pour le type 1, l'ordonnée sera: b_0, la pente: b_3
Pour le type 2, l'ordonnée sera: $b_0 + b_1$, la pente $b_2 + b_3$

Question 13: Comment feriez-vous si l'on avait 3 ou 4 types de roches pour effectuer le codage?

Remarque: Dans le dernier exemple, des résultats identiques (pour les coefficients) auraient été obtenus si l'on avait effectué les deux régressions séparément pour chaque type de roche.

2.4.6.1 Exemple: modélisation de la déformation du barrage de Beauharnois

Cet exemple est tiré du PFE de S. Lachambre et S. Dorion (1986). Le barrage de Beauharnois présente un important problème de déformation (expansion) en raison des réactions survenant entre les granulats (silice) et les alcalis du ciment. La réaction entraîne la formation d'un gel de silice accompagné d'une augmentation du volume du béton et de fissurations caractéristiques. Hydro-Québec a installé une série de repères sur le barrage dont la position est relevée précisément périodiquement (une mesure en été, une autre mesure en hiver). Avant de définir les variables, il faut noter que le barrage de Beauharnois a été construit en trois phases distinctes (1928, 1948 et 1956). Les étudiants ont cherché à établir un modèle permettant de décrire les déplacements observés en fonction des variables:

DEF: déformation (en mm) mesurée au repère (la variable Y de la régression). On dispose d'un total de 1158 mesures.

T3: température moyenne au cours des trois derniers jours précédant le relevé.

TM: température moyenne du mois ou la mesure a été effectuée.

STA: position géographique du repère le long du barrage.

JOUR: nombre de jours écoulés depuis le premier relevé.

P1,P2,P3: variables indicatrices; un repère pris sur la partie la plus ancienne du barrage aura P1=1, P2=0 et P3=0.

C1,C2,C3,C4: variables indicatrices prenant la valeur 0 si la mesure est effectuée avant la date de la coupure considérée et 1 après. Ces coupures sont des entailles effectuées à même le béton du barrage afin de permettre un relâchement des contraintes reliées au gonflement de l'ouvrage.

EVACU: une variable indicatrice pour identifier les repères se trouvant au-dessus d'un évacuateur de crues. Ces repères montrent un déplacement moindre en raison de la plus faible quantité de béton.

DECRO: une variable indicatrice pour identifier un décrochage (affaissement brusque survenu au repère 2 au relevé de février 1981).

Ces deux dernières variables indicatrices ont été introduites grâce à l'examen des résidus de la régression qui a mis en évidence le comportement très particulier de ces repères.

De plus, une multitude de variables additionnelles ont été formées en combinant certaines de celles-ci. Ainsi, les produits P1 JOUR, P2 JOUR, P3 JOUR permettent d'identifier des taux de déformations différents dans chaque partie du barrage. La variable C1 STA, permettrait de modeler l'effet de la coupure en tenant compte de la distance du repère par rapport à cette coupure. Le produit C1 STA JOUR permettrait en plus de tenir compte du facteur temps en relation avec cette coupure et en fonction de la distance par rapport à celle-ci.

Les auteurs obtiennent un R^2 de 0.91, en ne retenant que 6 variables grâce à une procédure "stepwise".

L'équation de régression obtenue est la suivante:

DEF=-4.2 -2.9 EVACU +.0035 JOUR -11.37 DECRO +.14 T3 -.0019 P3 JOUR +.014 TM

On remarque que:

- le barrage se déforme avec le temps
- la déformation est plus importante par temps chaud
- la partie nouvelle se déforme à un taux moindre que les deux plus anciennes parties
- la déformation est moindre aux évacuateurs de crues. Ceci semble confirmer l'hypothèse d'une déformation reliée au gonflement du béton.
- les coupures (C1 à C4) n'ont pas eu d'effet important puisqu'elles n'ont pas été retenues dans le modèle.

Finalement, l'examen des résidus (non présentés) montre que le modèle pourrait être encore amélioré. En effet, certains des relevés ont des résidus positifs pour toutes les stations, d'autres négatifs. Ceci indique qu'on pourrait chercher à améliorer la modélisation du temps ou de la température (par ex. ajouter des composantes quadratiques de JOUR ou de T3). Cependant, à R^2=0.91, le modèle est assez satisfaisant dans son ensemble.
Le barrage est subdivisé en trois parties distinctes. On pourrait vouloir déterminer si les trois parties du barrage se déforment à la même vitesse à partir d'un temps donné de référence. Si chaque section se déforme à la même vitesse, alors le modèle s'écrit:

$$DEF=B*JOUR$$

On compare ce modèle réduit au modèle complet suivant:

$$DEF=b_1*P_1*JOUR+b_2*P_2*JOUR+b_3*P_3*JOUR$$

Où P1,P2 et P3 sont des indicatrices (0 ou 1) servant à indiquer la section du barrage d'où provient la mesure. Si le barrage se déforme à la même vitesse dans chaque section, on devrait avoir $b_1=b_2=b_3=B$. Le test est identique à celui effectué pour l'ajout de variables. Ici le modèle réduit est le modèle avec un seul B, le modèle complet est celui avec b_1, b_2, b_3. Il n'y a pas de constante dans ce modèle, mais il pourrait y en avoir si on le désirait. La statistique à comparer à une $F_{2,n-3,1-\alpha}$ est

$$\frac{(SCE_r - SCE_c)/2}{SCE_c/(n-3)} \qquad 2.26$$

où l'indice "c" désigne le modèle complet et l'indice "r" le modèle réduit. Le nombre de degrés de libertés est donné au numérateur par la différence entre le nombre de paramètres dans chaque modèle. Au dénominateur, les d.l. sont donnés par le nombre d'observations moins le nombre total de paramètres dans le modèle complet. La validité de ce test provient comme toujours de l'orthogonalité entre les sommes de carrés présentes au numérateur et au dénominateur.

2.4.7 Exemples de régression et tests

On veut souvent vérifier si une équation de régression s'écarte d'un modèle théorique connu. Également, on peut vouloir vérifier si deux ou plusieurs ensembles de données fournissent les mêmes régressions.

2.4.7.1 Comparer une régression à un modèle théorique connu.

En hydrogéologie, il existe plusieurs relations empiriques permettant de prédire la perméabilité d'un dépôt meuble en fonction de paramètres tels la porosité, l'indice des vides ou la taille des grains. Une des plus utilisées est la relation de **Kozeny-Carman** qui est de la forme suivante:

$$k=C\ e^3/(1+e)$$

Dans sa maîtrise, Bussières (1993) a mesuré la perméabilité de divers résidus miniers. Il a cherché à établir, par régression, le lien entre indice des vides et perméabilité de façon expérimentale. La question s'est naturellement posée à savoir s'il obtenait une relation significativement différente de celle établie par Kozeny-Carman, i.e. peut-on appliquer l'équation générale de K-C au cas de résidus miniers.

Ses données pour la mine Solbec-Cupra sont les suivantes:

k (en cm/s)=1/1000*[0.1220 0.0389 0.3560 0.4110 0.1950 0.2580 0.2930 0.2410 0.5530 0.2440 0.0462 0.1300]
e= [0.71 0.58 0.78 0.82 0.77 0.74 0.78 0.69 0.87 0.72 0.67 0.78]

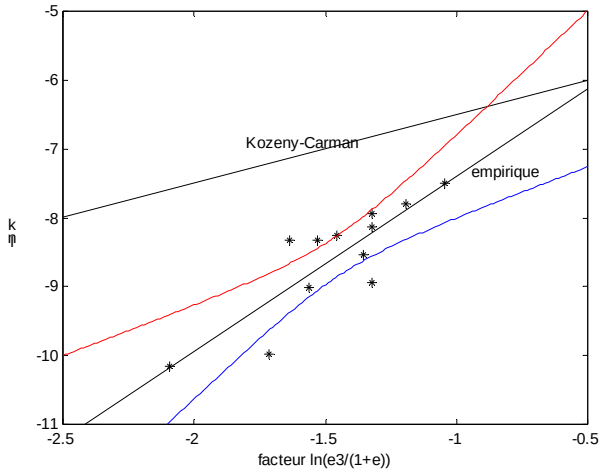
On a le modèle: $\ln(k)=b_0+b_1*\ln(e^3/(1+e))$

Utilisant les données de Bussière pour la mine Solbec-Cupra (n=12), on trouve:

b_0=-4.85 b_1=2.547 De plus, on trouve s_{b0}=.74 et s_{b1}=.50

Les coefficients du modèle de Kozeny-Carman sont B_0=-5.5, B_1=1. On trouve $t_{10,.95}$=2.28. L'intervalle de confiance autour de b_0 (i.e. ±2.28*.74) inclut la valeur B_0 du modèle K-C. Toutefois, l'intervalle pour b_1 (±2.28*.5) exclut la valeur B_1. Le modèle de Kozeiny-Carman n'est donc pas acceptable pour ce dépôt.

Une autre façon d'effectuer le test est de simplement tracer l'intervalle de confiance autour de la droite de régression et de vérifier si la droite de K-C s'y trouve incluse totalement.



2.4.7.2 Relation vent-vagues (Roy, 1995)

Dans sa maîtrise, N. Roy cherchait à prédire la hauteur et la période des vagues dans la rivière Outaouais, en fonction de la force du vent, afin d'effectuer le design de mesures de protection des berges. Des modèles ont été élaborés par d'autres chercheurs et sont présentés dans le "Shore Protection Manual". Roy voulait vérifier si ses observations correspondaient ou non aux formules théoriques existantes.

Ses données sont les suivantes:

ho,hc: hauteur de vague observée et calculée par la méthode SPM (cm).
to,tc: periode observée et calculée par la methode SPM (s).

ho= [1.6 3.3 4.6 2.2 6.7 7.5 7.7 8.5 10.7 10.9 12.2 11.7 11.0 13.9 14.2 14.9 15.0 15.9 17.4 19.6 21.1 21.7 22.6]
hc= [2.5 3.8 5.2 3.0 7.8 7.8 10.2 9.6 5.7 7.8 8.5 12.6 10.3 14.3 12.4 15.2 13.4 12.5 17.3 33.9 24.7 22.4 12.4]
to= [.8 .87 1 .77 .9 1.2 1.2 1.25 1.21 1.6 1.32 1.51 1.2 1.48 1.53 1.62 1.4 1.69 1.56 1.96 1.92 2. 1.9 2.16]
tc= [.74 .81 .96 .7 .96 1.1 1.21 1.09 .93 1.1 1.11 1.38 1.1 1.29 1.43 1.32 1.28 1.25 1.38 1.93 1.94 1.8 1.49,2.08]

Le modèle théorique ne comporte pas de constante, on impose donc également un modèle sans constante pour la régression. Ici les variables à expliquer sont ho et to à partir des observations sur le vent qui sont inclues dans le calcul de hc et tc. Les modèles sont donc:

$ho = b_h * hc + e$ et $to = b_t * tc + e$

Si les données observées sont compatibles avec le modèle SPM, alors les coefficients b_h et b_t devraient être voisins de 1. On trouve b_h=.919 et b_t=1.11. Les écarts-types sur ces coefficients sont respectivement de s_{bh}=.06 et s_{bt}=.02. La valeur seuil pour un intervalle de confiance de niveau 95% est $t_{23\text{-}1;.95}$=2.07. L'intervalle de confiance pour b_h inclut la valeur 1, mais non l'intervalle de confiance pour b_t. On conclut que la hauteur des vagues peut être estimée avec la formule théorique SPM, mais que la période doit être corrigée par le facteur b_t. Le modèle SPM a été élaboré pour une berge rectiligne infinie, ce qui n'est pas le cas d'une rivière. C'est un fait souvent observé que la période entre les vagues est sous-estimée par SPM pour les rivières.

2.4.7.3 Ajustement par moindres carrés pour la méthode de Cooper-Jacob

L'équation de Theis décrit le comportement de la surface piézométrique en fonction de la distance et du temps pour un piézomètre d'observation installé dans un aquifère confiné, homogène, infini et d'épaisseur constante. Cette équation est:

$$s = \frac{Q}{4\pi T}\left[-0.5772 - \ln(u) + u - \frac{u^2}{2 \bullet 2!} + \frac{u^3}{3 \bullet 3!} + \ldots\right]$$

où s est le rabattement

Q est le débit pompé
T est la transmissivité
$u = r^2 S / 4Tt$
r est la distance au puits du piézomètre d'observation
S est le coefficient d'emmagasinement
t est le temps

Les inconnues sont S et T que l'on doit déterminer à partir de s,t, r et Q qui sont observés. L'équation de Theis est non-linéaire et pourrait être solutionnée comme telle (voir section 2.8). Toutefois, pour de faibles valeurs de u, les deux premiers termes entre crochets sont prépondérants. Cooper et Jacob ont utilisé cette propriété pour développer leur méthode graphique. Ici, on va utiliser une régression linéaire pour estimer S et T.

On utilise le modèle suivant:

$$s = b_0 + b_1 \ln(t) + e$$

À partir de b_0 et b_1 et comparant avec l'équation de Theis, on trouve les relations suivantes:

$$T = \frac{Q}{4\pi b_1}$$

$$S = \frac{2.25\,T}{r^2 e^{b_0/b_1}}$$

exemple: les données suivantes viennent de Todd (p. 127).

t(jour)= [1 1.5 2 2.5 3 4 5 6 8 10 12 14 18 24 30 40 50 60 80 100 120 150 180 210 240]/(60*24);
s(m)= [.2 .27 .3 .34 .37 .41 .45 .48 .53 .57 .6 .63 .67 .72 .76 .81 .85 .9 .93 .96 1 1.04 1.07 1.10 1.12];

r=60 m Q=2500 m^3/j

On trouve b_0=1.422 b_1=.1703 (R^2=.9992)

d'où: T=1168 m^2/j S=0.00017

Todd, trouve par méthodes graphiques:

Theis:	T=1110	S=0.00021
Cooper-Jacob:	T=1160	S=0.00018
Chow:	T=1160	S=0.00021

Ces valeurs sont très semblables à celles obtenues par régression.

2.4.7.4 Coefficients de récession d'aquifères et de bassins hydrographiques

Dans un TP du cours d'hydrogéologie on doit, à partir des débits mesurés dans une rivière après une période de crue, calculer les coefficients de récession total et de l'aquifère. Le coefficient de récession est simplement le taux de décroissance du débit en fonction du temps. On l'estime à partir d'un graphe où l'on porte en y le débit au jour j+1 et en x le débit au jour j.

Les données sont:

t(j)= [1 2 3 ...13]
Q(m^3/j)= [972 708 397 254 163 122 92 78 68 58 50 43 37]

Au début de la récession, l'eau de ruissellement, en plus de l'eau de l'aquifère contribue au débit enregistré. Après un certain temps, le ruissellement cesse et l'eau n'est fournie que par l'aquifère. Au début, la diminution du débit est très rapide, après, elle est plus lente. Au début, on parle de récession totale, à la fin, de récession de l'aquifère. Il faut déterminer à quel moment le ruissellement cesse. Ceci est habituellement fait visuellement à partir d'un graphe Q_{j+1} vs Q_j. Nous voyons ici une façon simple par régression de déterminer le moment où le ruissellement cesse et d'estimer les deux coefficients de récession (total et aquifère).

Il est utile de procéder d'abord à un examen sur échelle log-log pour identifier les points pouvant servir à l'ajustement des deux droites. Pour chaque partie linéaire, on ajuste une droite:

$Q_{j+1} = r_a Q_j$ j après la période de crue
$Q_{j+1} = r_t Q_j$ j en période de crue avant la transition

Noter que le modèle ne contient pas de constante. Effectivement, si la rivière est à sec au jour j, elle le sera au jour j+1. Après examen des données, on a retenu les points j=2, 3 et 4 pour évaluer r_t et 7 à 12 pour évaluer r_a. On a trouvé r_a=0.86 et r_t=0.59. On a effectué la régression avec les valeurs de débit, bien qu'on aurait pu aussi les réaliser sur les log(debits), auquel cas le problème revient à trouver la moyenne de $\log(Q_{j+1})-\log(Q_j)$. De cette façon, on trouve r_a=.86 et r_t=.61.

2.4.7.5 Détermination des vitesses (lenteurs) sismiques.

Pour simplifier, considérons un domaine découpé en « n » blocs réguliers chacun ayant une vitesse sismique inconnue vi et supposons que le trajet entre la source et le récepteur suit une ligne droite (ceci est physiquement faux car on sait, par la loi de Snell, qu'un rayon sera réfracté à l'interface de deux milieux de vitesses différentes; toutefois cette approximation peut être acceptable en première approximation si les contrastes de vitesses ne sont pas trop élevés).

Le temps de parcours de l'onde sismique le long d'un rayon « i » dans un bloc « j » est donné par :

$t_{ij}=l_{ij}/v_j$

où t_{ij} est le temps de parcours de l'onde, l_{ij} est la longueur du rayon « i » dans le bloc « j » et v_j est la vitesse dans le bloc j.

En introduisant le changement de variable $w_j=1/v_j$, on a :

$t_{ij}=l_{ij}*w_j$

En supposant des rais droits on peut exprimer le temps de parcours le long de chaque rai comme une sommation des lenteurs de chaque bloc x longueur parcourue par le rayon dans chaque bloc (ce n'est qu'une approximation car on sait que les rayons sont réfractés aux interfaces de blocs de vitesses différentes). On obtient ainsi une équation linéaire de type régression. Les t_{ij} sont les valeurs observées (le Y de la régression), les l_{ij} sont les paramètres fixés (la matrice X de la régression) puisqu'ils ne dépendent que de la configuration géométrique émetteur-récepteur et de la discrétisation en blocs qui est effectuée. Finalement les w_j sont les inconnues du système (les coefficients « b » de la régression).

Exemple (inspiré de Tarantola[1], 1987, p.92):

s1	s2	s3
1	4	7
2	5	8
3	6	9

r1 r2 r3 r4

On désire déterminer les « lenteurs » (w_j) des cellules 1 à 9 à partir d'émissions (s1, s2 et s3) situées au centre des faces supérieures et des récepteurs r1 à r4 situés aux limites des blocs sur leur face inférieure.

On aura 12 temps de parcours enregistrés (12 rayons) et on a 9 coefficients à estimer (les lenteurs des 9 blocs). Le vecteur y est donc ici 12x1 et la matrice X 12x9, les coefficients b 9x1. Le problème de régression ici est sans constante.

Si l'on effectue la régression telle quelle, on obtient une erreur! La raison est que la matrice X'X qui est d'ordre 9 a un rang de 7 et est donc singulière. Cette singularité vient des relations existant entre les différents rayons (lignes de X). En effet, on a :

s1r1=s1r2 (où s1r2 désigne le temps de parcours de la source s1 au récepteur r2)
s2r2=s2r3
s3r3=s3r4
c*(s1r2+s2r2)-(s1r3+s2r1)=0 avec c=1.10282....=cos(atan(0.5/3))/cos(atan(1.5/3))
etc.

Solutions :

Différentes approches peuvent être utilisées :

- Changer la position et/ou la taille et la forme des blocs découpant le volume afin de diminuer la symétrie du problème et/ou le nombre de paramètres à estimer.
- Ajouter des données par exemple en plaçant de nouvelles sources du côté gauche du bloc et des récepteurs à droite (si c'est possible).
- Imposer des contraintes au modèle afin de réduire le nombre d'inconnues (par exemple fixer la lenteur sismique de deux des 9 blocs).
- Parmi l'infinité de solutions possibles se donner un critère statistique permettant d'en choisir une. Cette approche porte le nom d'inversion (pour les systèmes sous-déterminés). La régression n'est rien d'autre en fait qu'une inversion (sur-déterminée). Cette approche consiste essentiellement à remplacer la matrice inverse $(X'X)^{-1}$, qui n'existe pas, par son inverse généralisée $(X'X)^{-g}$ (on peut aussi simplement prendre l'inverse généralisée de X et poser $b=X^{-g}$ y. Les deux approches donnent des résultats identiques).

Note : Une inverse généralisée, que l'on appelle parfois pseudo-inverse ou inverse de Moore-Penrose est telle que pour M une matrice, pouvant être rectangulaire,

on a : $M\,M^{-g}\,M = M$
et $M^{-g}\,M\,M^{-g} = M^{-g}$

On peut construire une inverse généralisée de M à partir de sa décomposition en valeurs singulières (SVD). (Voir Tarantola, 1987; Menke[2], 1984).

[1] Tarantola, 1987, Inverse problem theory, Elsevier.
[2] Menke, 1984, Geophysical data analysis, discrete inverse theory

2.4.7.6 Détermination de la densité de blocs à partir des anomalies gravimétriques mesurées en surface (et éventuellement en forage).

L'anomalie gravimétrique en un point (x_0, y_0, z_0), due à un corps Ω avec densité $\rho(x,y,z)$ est donnée par :

$$g(x_0, y_0, z_0) = -G\iiint_{\Omega} \rho(x,y,z) \frac{z_0 - z}{\left\{(z_0-z)^2 + (y_0-y)^2 + (x_0-x)^2\right\}^{3/2}} dxdydz$$

où G est la constante d'attraction gravitationnelle universelle.

Si Ω est un parallélépipède ayant une densité constante, on peut alors simplifier l'intégrale triple précédente à une triple sommation sur les sommets du parallélépipède (voir Plouff[3](1976)) :

$$g(x_0, y_0, z_0) = -G\rho\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\mu_{ijk}\left[\Delta x_i \ln(\Delta y_j + r_{ijk}) + \Delta y_j \ln(\Delta x_i + r_{ijk}) - \Delta z_k \arctan\left(\frac{\Delta x_i \Delta y_j}{\Delta z_k r_{ijk}}\right)\right] \quad (2)$$

où:

r_{ijk} est la distance entre le point (x_0,y_0,z_0) et le sommet du parallélépipède identifié par les indices ijk,
$\mu_{ijk} = (-1)^{(i+j+k)}$
$\Delta x_i = x_0 - x_i$, $\Delta y_j = y_0 - y_j$, and $\Delta z_k = z_0 - z_k$
ρ est la densité constante du bloc.

Bref, (2) est de la forme $g_0 = c_{0i}\,\rho_i$. Si maintenant on considère plusieurs points pour lesquels on mesure l'anomalie gravimétrique et plusieurs blocs représentant le domaine, on peut construire la relation :

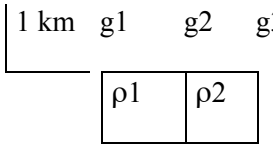
$$g = C\rho + e$$

où « g » est un vecteur nx1 (les n points où l'anomalie gravimétrique a été mesurée)
« C » est la matrice géométrique n x m. Les coefficients de cette matrice s'obtiennent en évaluant (2). Ils sont uniquement fonction de la position relative du point et des sommets du bloc.
« ρ » est le vecteur mx1 des densités inconnues que l'on cherche à estimer à partir des mesures faites de gravimétrie.
« e » est le vecteur d'erreurs de mesure

Si on a n>m, alors on a un contexte de régression. Ici la matrice C joue le rôle de la matrice X dans les sections précédentes, le vecteur « g » est le « y » de la régression et le vecteur « ρ » est le « b » de la régression.

Si on a n<m, alors (C'C) sera singulière et il y aura donc une infinité de solutions possibles. Dans ce cas, en pratique on adoptera une solution ayant certaines propriétés de régularité (ex. longueur minimale du vecteur ρ). Ceci conduit à utiliser encore une fois l'inverse généralisée de C (ou de C'*C) pour estimer ρ.

Exemple numérique : Soit la disposition suivante :



[3] Plouff, 1976, Gravity and magnetic field of polygonal prisms and application to magnetic terrain correction

On calcule (avec (2)) C=

$$\begin{bmatrix} 969.3881 & 108.1597 \\ 969.3881 & 969.3881 \\ 108.1597 & 969.3881 \end{bmatrix}$$

Si l'on pose $\rho=[0\ 0.2]'$

On calcule $g=[\ 21.6319,\ \ 193.8776,\ \ 193.8776]'$

Partant cette fois de g et estimant ρ, on trouve comme il se doit :

$\rho est= (C'*C)^{-1}*C'*g = [0,0.2]'$

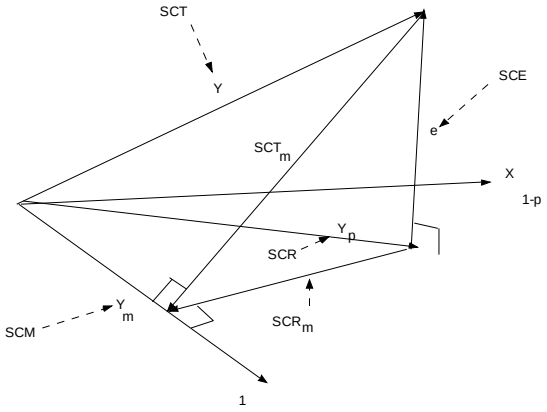
Si on ajoute une erreur aux données de g, alors ρest sera la solution moindre carrés. Ainsi, supposons que l'on a observé plutôt : g=[21 194 193]', on calcule alors

ρest= [-.0001 0.1996]

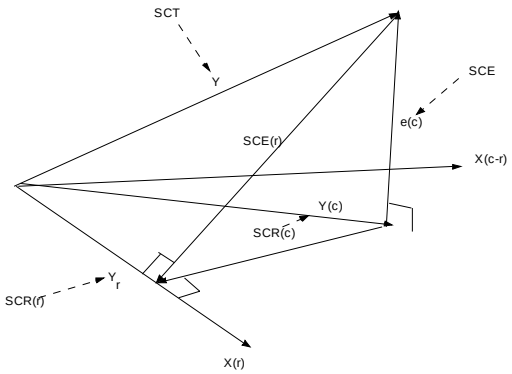
2.5 Géométrie des moindres carrés

La figure suivante représente en termes géométriques la régression. On voit que la régression n'est rien d'autre que la projection du vecteur Y dans l'espace engendré par les colonnes de X. De ceci découle les orthogonalités décrites précédemment.

Géométrie de la régression, modele avec constante



Géométrie de la régression, modele réduit (r) vs complet (c)



2.6 Corrélation partielle

Dans le calcul de corrélations simples, tous les facteurs sont confondus. Très souvent on est intéressé à éliminer l'effet (linéaire) d'une ou de plusieurs variables avant de calculer les corrélations entre les variables qui nous intéressent. C'est ce que l'on effectue en calculant les corrélations partielles.

Supposons que l'on ait trois variables X, Y et Z. On pourrait vouloir calculer la corrélation entre Z et Y après avoir éliminé l'effet linéaire de X sur ces deux variables. Pour éliminer l'effet linéaire de X, on n'a qu'à effectuer la régression de Z sur X et conserver les résidus. Ceux-ci représentent la part de Z qui ne peut être linéairement expliqué par X. On peut faire de même en régressant Y sur X et en conservant les résidus de cette régression. La corrélation simple entre ces deux ensembles de résidus est appelée corrélation partielle de Z avec Y étant donné l'effet linéaire de X filtré.

Exemple: On sait que le TiO_2 et le SiO_2 sont de bons indices de la maturité magmatique des roches volcaniques. On pourrait vouloir éliminer l'effet de la différenciation magmatique sur les corrélations entre les autres variables. Lors de la différenciation magmatique, les minéraux ferro-magnésiens cristallisent en premier. On observera donc typiquement une corrélation positive entre FeO et MgO. Cependant, ces deux éléments se trouvent en compétition pour occuper les mêmes sites de cristallisation sur les minéraux. Ceci entraîne que pour des roches de maturité magmatique comparable, on devrait observer une corrélation négative entre FeO et MgO. C'est ce que nous permettrait de voir les corrélations partielles.

Soit X, Y et Z trois variables dont on a soustrait la moyenne (i.e. elles sont centrées). La corrélation partielle entre Y et Z (étant donné X) est la corrélation simple entre Y et Z étant donné l'effet linéaire de X enlevé.

Soit: $Y_{.x} = Y - a\,X$
$Z_{.x} = Z - b\,X$

où a et b sont les coefficients obtenus de la régression.

$a = s_{xy} / s_x^2$
$b = s_{xz} / s_x^2$

La corrélation (simple) entre $Y_{.x}$ et $Z_{.x}$ s'écrit :

$s_{Y.x\,Z.x} / (s_{Y.x}\, s_{Z.x})$

On calcule:

$s_{Y.x\,Z.x} = s_{yz} - a\, s_{xz} - b\, s_{xy} + ab\, s_x^2$

$s_{Y.x}^2 = s_y^2 - 2a\, s_{xy} + a^2\, s_x^2$

$s_{Z.x}^2 = s_z^2 - 2b\, s_{xz} + b^2\, s_x^2$

Substituant les valeurs pour a et b et simplifiant, on arrive à:

$$r_{yz\bullet x} = \frac{r_{yz} - r_{xy} r_{xz}}{\sqrt{\left[(1 - r_{xy}^2)\,(1 - r_{xz}^2)\right]}} \qquad 2.27$$

Remarques:

i. On peut calculer toutes les corrélations partielles à partir de la matrice des corrélations simples.

ii. On pourra enlever l'effet linéaire d'une deuxième variable, puis d'une troisième ... de façon récursive i.e. on applique la formule ci-dessus en remplaçant dans la formule les corrélations simples par les corrélations partielles de l'étape précédente.

ex. Supposons que l'on veut maintenant éliminer l'effet linéaire de W en plus de X. On calculera:

$$r_{yz\bullet xw} = \frac{r_{yz\bullet x} - r_{wy\bullet x} r_{wz\bullet x}}{\sqrt{\left[(1 - r^2_{wy\bullet x})\,(1 - r^2_{wz\bullet x})\right]}} \qquad 2.28$$

iii. On peut démontrer que l'ordre dans lequel on "fixe" les variables n'influence pas les résultats. Pour ce faire on n'a qu'à effectuer les calculs en fixant X puis W et recommencer en fixant cette fois W puis X.

iv. La dérivation de la corrélation partielle suggère qu'il peut être instructif d'examiner le diagramme binaire des résidus dans un problème de régression. Ainsi, étant donné les variables X déjà dans la régression, on peut calculer les résidus pour Y et les résidus pour les autres X non inclus dans la régression (i.e. chaque X est régressé par les mêmes X que Y). Une variable ne peut entrer dans la régression que si ce diagramme (résidu Y vs résidu X) montre une relation linéaire suffisamment forte.

Exemple numérique: On a mesuré SiO_2, MgO et FeO et on a obtenu, avec 30 observations, les corrélations simples suivantes entre ces trois éléments:

	SiO_2	MgO	FeO
SiO_2	1	-0.86	-0.75
MgO		1	0.50
FeO			1

La corrélation partielle entre le MgO et le FeO (étant donné l'effet linéaire de SiO_2 enlevé) est :

$[0.5-(-0.86)(-0.75)] / [(1- 0.86^2)(1-0.75^2)]^{0.5} = -0.43$

La situation s'est complètement inversée par rapport au coefficient de corrélation simple!

Sous une forme plus générale, soit p variables pour lesquelles on veut calculer les corrélations partielles entre les q premières variables étant donné l'effet linéaire des variables q+1, q+2...p enlevé. On définit une matrice de covariances partielles (i.e. une matrice de covariances pour lesquelles on a éliminé l'effet d'autres variables) à partir de laquelle on pourra construire les corrélations partielles exactement comme on le fait pour les corrélations simples.

Soit:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad 2.29$$

où Σ_{11} est la matrice q x q des covariances entre les q variables d'intérêt.
$\Sigma_{12} = \Sigma_{21}'$ est la matrice q x (p-q) des covariances entre les variables d'intérêt et les variables dont on veut éliminer l'effet linéaire.
Σ_{22} est la matrice (p-q) x (p-q) des covariances entre les p-q variables que l'on veut fixer.

La matrice des covariances partielles s'écrit :

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}$$

Il est relativement facile de vérifier que les corrélations partielles peuvent s'obtenir directement de cette matrice en prenant les covariances et les variances requises.

Exercice 6: Avec p=3 et q=2 vérifiez que l'on obtient le même résultat avec la formule récursive qu'avec la matrice de covariances partielles.

2.6.1 Lien entre corrélation partielle et régression

Ces deux notions sont extrêmement liées:

i. Le coefficient d'une variable dans une régression multiple peut s'obtenir de la matrice de covariance partielle (cf. cas de deux variables).
ii. Lorsqu'on ajoute une variable à un modèle existant, on a la relation suivante entre les coefficients de corrélation multiple et le coefficient de corrélation partielle:

$$\frac{R_p^2 - R_{p-1}^2}{1 - R_{p-1}^2} = r^2_{yx_p|x_1,x_2\cdots x_{p-1}} \qquad 2.30$$

Le carré de la corrélation partielle, donne donc l'augmentation de R^2 relative à la portion de la variation de y inexpliquée par les variables déjà dans l'équation.

Exercice 7: En vous servant du test F développé précédemment pour tester le caractère significatif d'un ajout de variables, développez un test pour vérifier le caractère significatif de la corrélation partielle.

Question 14: Dites comment la corrélation partielle peut être utile dans une procédure de type « stepwise ». Y aurait-il des avantages du point de vue temps de calcul?

2.7 Tests sur les coefficients de corrélations simples et partielles

En plus des tests vus précédemment, on peut tester une hypothèse beaucoup plus générale à l'aide du test suivant (test de Fisher; valide pour n>25):

$H_0 : \rho_{xy} = \rho_0$
$H_1 : \rho_{xy} \neq \rho_0$

Sous H_0, on peut montrer que :

$$\frac{1}{2}\ln\left[\frac{(1+r_{xy})}{(1-r_{xy})}\right] \approx N\left(\frac{1}{2}\ln\left[\frac{(1+\rho_0)}{(1-\rho_0)}\right], \frac{1}{(n-3-nfix)}\right) \qquad 2.31$$

où nfix indique le nombre de variables fixées s'il s'agit d'une corrélation partielle (=0 si c'est une corrélation simple).

Remarque: Ce test est plus général que le test déduit de la régression car ce dernier ne permet de vérifier que si la corrélation est significative, i.e. différente de zéro. Ici on peut tester si la corrélation est significativement différente de n'importe qu'elle corrélation ρ_0 déterminée à priori.

Exemple: Avec 30 observations, on a obtenu r_{xy}=0.4. Est-ce significativement différent de 0.5?

On calcule 0.5 ln(1.5/0.5) = 0.55
0.5 ln(1.4/0.6) = 0.42
et (0.42-0.55) $\sqrt{27}$ = -0.68

Comparant cette valeur avec la valeur critique tirée d'une table de la loi normale (-1.96 avec α=0.05, test bilatéral), on conclut que l'on doit accepter H_0, i.e. on ne peut rejeter l'hypothèse que le vrai coefficient de corrélation de la population soit égal à 0.5 .

Exemple: La corrélation partielle, obtenue avec 30 observations, entre MgO et FeO, étant donné l'effet linéaire de SiO_2 éliminé (on avait trouvé précédemment $r_{FeO\ MgO\ .\ SiO2}$= -0.43), est-elle significativement différente de 0.5?

On calcule 0.5 ln (1.5/0.5) = 0.55
0.5 ln (0.57/1.43) = -0.46
et (-0.46-0.55)$\sqrt{26}$ = -5.15

On rejette fortement l'hypothèse H_0 (valeur critique =-1.96 au niveau 0.05 , bilatéral). On peut donc conclure que le SiO_2 a un effet significatif sur la corrélation entre MgO et FeO.

Cet exemple illustre bien qu'un énoncé concernant les corrélations entre variables n'a de valeur réelle que lorsque les conditions expérimentales sont clairement énoncées et que les facteurs extérieurs sont le plus possible pris en considération. Les corrélations ne font que décrire des liens linéaires entre variables et aucune conclusion ne peut être énoncée face à un éventuel lien causal entre variables.

Question 15: Dans une séquence sédimentaire, vous mesurez l'épaisseur de carbonates et de schistes. Vous convertissez vos épaisseurs brutes en épaisseurs relatives (i.e. proportions). Quelle corrélation observerez-vous entre les deux épaisseurs relatives?

Question 16: Dans le même contexte, vous calculez la corrélation partielle entre les épaisseurs brutes de shales et de carbonates, étant donné la variable "épaisseur totale" fixée. Quelle corrélation partielle obtiendrez-vous?

Question 17: Cherchant à extrapoler ce qui précède, les corrélations entre variables présentant une fermeture (i.e. leur somme est constante) auront-ils tendance à montrer plus de corrélations positives ou négatives?

2.8 Exemple numérique complet

Soit y = [6 4 20 24]
x1= [5 10 15 20]
x2= [1 1 2 2]

On forme la matrice X

1	5	1
1	10	1
1	15	2
1	20	2

On trouve X'X=

4	50	6
50	750	85
6	85	10

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

2.75	0.1	-2.5
-0.1	.04	-.4
-2.5	-.4	5

et X'Y = [54 850 98]'

b = [-11.5 0.2 15]

yc = [4.5 5.5 21.5 22.5]'

e=y-yc = [1.5 -1.5 -1.5 1.5]'

	SC	CM	dl
SCT	1028	257	4
SCR	1019	340	3
SCE	9	9	1
SCM	729	729	1
SCT_m	299	100	3
SCR_m	290	145	2

r2 = 0.9699

La matrice de variance-covariance des coefficients b est:

b_0	24.75	0.9	-22.5

b_1	0.9	0.36	-3.6
b_2	-22.5	-3.6	45.

2.9 Complément sur les régressions

Nous abordons ici, pêle-mêle et très brièvement, certains sujets qui seraient normalement vus dans un cours plus approfondi sur les régressions.

2.9.1 Régressions non-linéaires

Lorsque la régression ne peut, par transformations, être linéarisée, alors la seule méthode de solution est par itération. On entre alors dans le domaine de l'optimisation d'une fonction « objectif » avec toutes les difficultés que cela peut impliquer (non-convergence, convergence vers des optimums locaux, calculs importants). Ces problèmes augmentent avec le nombre de paramètres à estimer.

Une autre méthode courante consiste à utiliser l'expansion en série de Taylor de la fonction et d'appliquer la régression à la partie linéaire évaluée en un point initial pas trop éloigné de la solution cherchée. On trouve alors un 2^e point où l'on évalue à nouveau l'expansion de Taylor et ainsi de suite jusqu'à l'optimum.

Ainsi supposons que l'on a :

Y=f(b) et f(b) est une fonction non-linéaire. On effectue l'expansion de Taylor autour d'un point b^0 où l'exposant indique le numéro de l'itération.

$f(b) \approx f(b^0)+df(b^0)/d(b)\ (b^1-b^0)$

b^0 étant connu, la seule inconnue est b^1 et on a un problème de régression linéaire en b^1. On estime donc b^1 puis on écrit :

$f(b) \approx f(b^1)+df(b^1)/d(b)\ (b^2-b^1)$

et on estime b^2 et ainsi de suite jusqu'à ce que l'on obtienne convergence. Plus généralement, à l'itération k+1, on obtiendra les nouveaux coefficients b^{k+1} à partir des coefficients à l'étape b^k par l'équation de régression:

$$Y = \hat{Y}^k + J^k (b^{k+1} - b^k) + e$$

i.e.

$$Y = (\hat{Y}^k - J^k b^k) + J^k b^{k+1} + e$$

où

$$\hat{Y}^k = f(b^k)$$

$$b^{k+1} = \left(J^{kT} J^k\right)^{-1} J^{kT} \left(Y - \hat{Y}^k + J^k b^k\right) = b^k + \left(J^{kT} J^k\right)^{-1} J^{kT} \left(Y - \hat{Y}^k\right)$$

Cette dernière équation indique que les coefficients à l'itération k+1 sont égaux aux coefficients de l'itération k auxquels on ajoute une perturbation donnée par la régression des résidus de l'étape précédente.

On a alors un problème de régression linéaire pour les coefficients b^{k+1} (la matrice J^k est la matrice des dérivées premières de la fonction, la jacobienne, estimées aux valeurs b^k). La matrice J^k a pour dimension n x p où n est le nombre d'observations et p est le nombre de paramètres à estimer. Dans le cas d'un modèle linéaire, la matrice J n'est rien d'autre que la matrice X de la régression.

L'algorithme implique donc 3 étapes principales:

i. Évaluer $f(b^k)$
ii. Évaluer J^k
iii. Évaluer b^{k+1}

Après l'étape iii, on compare les nouveaux b^{k+1} aux anciens b^k, s'ils n'ont pas changé significativement, on arrête.
ex. Supposons que l'on ait $Y=b_0+x_1^{b1}$. On cherche à estimer b_0 et b_1

Soit Y=[3 4.8284 7.1962 10 13.1803]'
et x_1=[1 2 3 4 5]'

On vérifiera que la solution est b_0=2 et b_1=1.5.

Supposons que l'on ait comme solution initiale b_0^0=1, b_1^0=1.
On a alors : $f(b^0)$=1+[1 2 3 4 5] ^{1} = [2 3 4 5 6]

L'expansion de Taylor peut s'écrire :

$Y=f(b)\approx f(b^k)+df(b^k)/d(b)\ (b^{k+1}-b^k)$

= [2 3 4 5 6]'+ J^0 [b_0^1-1 b_1^1-1]'

J^0=[1 x_i^{b1} ln(x_i)] i=1...5
J^0=[1 0
1 1.3863
1 3.2958
1 5.5452
1 8.0472]

Par régression, on trouve à la 1ère itération:
b_o^1=1.8193 b_1^1=1.772.

$f(b^1)$=1.8193+[1 2 3 4 5] $^{1.772}$= [2.8193 5.2346 8.8251 13.4834 19.1404]
J^1=[1 0
1 2.3673
1 7.6967
1 16.17
1 27.877]

On trouve alors :
b_0^2=1.9360 b_1^2=1.5524

et ainsi de suite. À la 5e itération, on trouve :

b_0^5=2 et b_1^5=1.5 qui est la solution exacte à ce petit problème.

2.9.2 Régression logistique[4]

Dans le cas particulier où la variable Y observée est une variable binaire (0 ou 1, succès ou échec), le modèle de régression linéaire ne s'applique pas vraiment. Ce cas est rencontré très fréquemment en pratique et plusieurs méthodes ont été développées pour l'étudier. Une de ces méthodes est l'analyse discriminante (voir chapitre 5). Une autre méthode est la régression logistique.

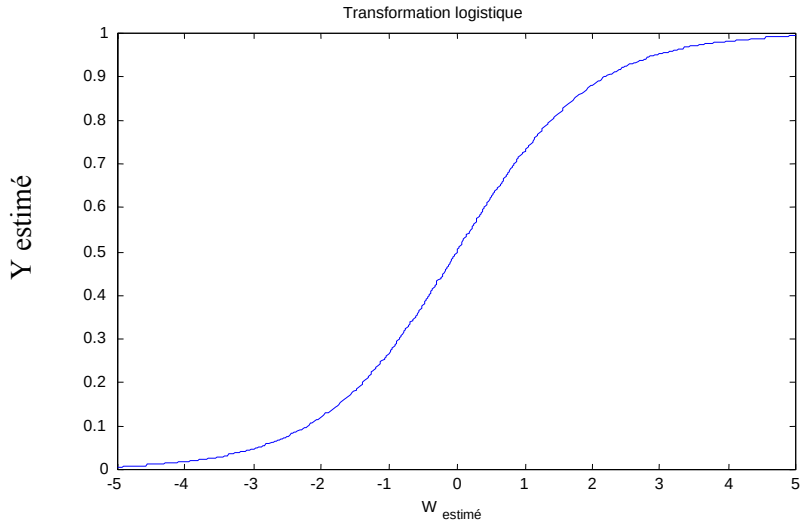
Interprétons Y comme une probabilité. La valeur 1 représente le fait que l'on est certain que l'événement se réalise, la valeur 0 que l'on est certain qu'il ne se réalise pas. Toute valeur comprise entre 0 et 1 décrit la probabilité que l'événement se réalise. L'idée est d'utiliser une transformation symétrique de cette probabilité qui associe la valeur 1 à l'infini, 0 à moins l'infini et 0.5 à 0. La transformation logistique effectue précisément cela. On la définit comme:

W=log(Y/(1-Y)). Si Y=1, W=infini; si Y=0, W=-infini; si Y=0.5, W=0 et on a W(Y)= -W(1-Y) (symétrie).

On effectue la régression de W sur les variables explicatives, discrètes ou continues, contenues dans X.
Le modèle de prédiction est donc:

$$\hat{W} = b_0 + b_1 X_1 + ... + b_p X_p$$

La figure suivante montre la relation existant entre $\hat{W}$ et $\hat{Y}$:



On ne peut toutefois estimer les coefficients "b" en minimisant la somme des carrés des erreurs puisque les valeurs "observées" de W sont + ou – infini (Y vaut 0 ou 1). On utilise plutôt la méthode de vraisemblance maximale. La vraisemblance est la fonction de probabilité conjointe évaluée selon le modèle de régression aux valeurs observées de Y. La vraisemblance maximale est la vraisemblance la plus forte que l'on peut obtenir parmi tous les choix possibles de régression (et donc de $\hat{W}$ et $\hat{Y}$).

Pour un vecteur X donné, la régression prédit une probabilité $\hat{Y}$ d'être 1 et une probabilité (1- $\hat{Y}$) d'être 0. Si l'on a observé Y=1, la vraisemblance de l'observation est $\hat{Y}$. Si l'on a observé 0, la vraisemblance de l'observation est (1- $\hat{Y}$). Pour 2 observations, sous hypothèse d'indépendance, la vraisemblance du couple (0,0) est $(1-\hat{Y}_1)(1-\hat{Y}_2)$; la vraisemblance du couple (0,1) est $(1-\hat{Y}_1)\hat{Y}_2$; la vraisemblance du couple (1,0) est

4 Référence principale: Hosmer et Lemeshow (1989), Applied logistic regression, Wiley.

$\hat{Y}_1(1-\hat{Y}_2)$; la vraisemblance du couple (1,1) est $\hat{Y}_1\,\hat{Y}_2$. Considérant conjointement toutes les observations, sous hypothèse d'indépendance

$$L = \prod_{i,Y_i=1} \hat{Y}_i \prod_{j,Y_j=0} (1-\hat{Y}_j)$$

$$\text{où } \hat{Y} = \frac{exp(\hat{W})}{1+exp(\hat{W})}$$

Note : $\hat{Y}$ est compris entre 0 et 1. La valeur maximale théorique de la vraisemblance est donc 1 qui se produit si tous les $\hat{Y}$ sont 1 quand Y est 1 et sont 0 quand Y est 0. La régression logistique cherche la combinaison des valeurs observées X qui fournit la plus grande séparation possible entre les observations des 2 groupes définis par Y=0 et Y=1.

Note : On peut généraliser la régression logistique au cas où plus de 2 groupes (k groupes) sont présents. Il suffit d'identifier 1 des groupes comme groupe référence et de traiter tous les autres groupes en succession relativement à ce groupe. On obtient ainsi k-1 équations de régression et l'on peut estimer la probabilité qu'une observation donnée appartienne à un groupe particulier. Le choix du groupe de référence n'affecte aucunement les probabilités calculées.

En pratique, on maximisera plutôt log(L). Cette maximisation nécessite d'utiliser des procédures itératives (régression non-linéaire). La méthode la plus courante utilise une régression pondérée où la matrice de pondération est recalculée à chaque itération en fonction de "b" obtenus à l'itération précédente.

Algorithme:

i. Spécifier b_0, poser $b=b_0$

ii. Calculer $\hat{W}=Xb$

iii. Calculer $\hat{Y}=\frac{\exp(\hat{W})}{1+\exp(\hat{W})}$

iv. Calculer la matrice de variance: $V=\text{diag}(\hat{Y}(1-\hat{Y}))$. V est diagonale de taille n x n. (Note: variance d'une loi Bernouilli: (p*(1-p))

v. Calculer les coefficients mis à jour $b_{k+1}=b_k+(X'VX)^{-1}X'(Y-\hat{Y})$

vi. Poser $b=b_{k+1}$, aller à ii (jusqu'à convergence).

vii. La matrice de variance-covariance des coefficients de la régression est $V(b)=(X'VX)^{-1}$.

La régression logistique est très couramment utilisée en médecine et en pharmacologie et les études épidémiologiques où l'on a naturellement une variable Y de type dichotomique : effet ou pas d'effet d'un médicament ou d'une procédure en fonction su sexe, de l'âge, du poids, du milieu de vie, etc.; mort ou survie des patients en fonction de tel traitement,...

$\hat{Y}/(1-\hat{Y})$ est ce que l'on appelle en anglais un "odds ratio". $\hat{W}$ est donc le log(odds ratio). Les coefficients de la régression indiquent pour chaque incrément de la variable "x" considérée l'accroissement de ce log(odds ratio). Donc exp(b) exprime le facteur multiplicatif d'accroissement du odd-ratio fourni par l'accroissement d'une unité de la variable x.

Note : L'équation présentée au point v. reprend les résultats de la section précédente sur la régression non-linéaire avec la modification suivante. Comme les résidus de la régression sont soit $\hat{Y}$ (si Y=0) soit (1- $\hat{Y}$) (si Y=1), la variance des résidus est $\hat{Y}$ (1- $\hat{Y}$). Il est logique alors de donner plus de poids aux résidus de faible variance qu'aux résidus à forte variance. Si l'on place ces variances dans une matrice diagonale V, ceci revient à minimiser SCE=e'V^{-1}e plutôt que e'e comme on le fait en régression. On parle alors de régression pondérée. Dans une régression linéaire pondérée, on a b=(X'V^{-1}X)$^{-1}$X'V^{-1}Y où X est la matrice dans le modèle Y=Xb+e. Dans le cas qui nous occupe, le rôle de X est joué par la matrice jacobienne. Or $J = \frac{d\hat{Y}}{db} = \frac{d}{db}\frac{\exp(Xb)}{(1+\exp(Xb))} = XV$ où X est la matrice dans le modèle $\hat{W} = Xb$. L'équation non-linéaire de régression avec pondération sur les résidus est donc :

$$b^{k+1} = b^k + \left(J^{kT}V^{-1}J^k\right)^{-1}J^{kT}V^{-1}\left(Y - \hat{Y}^k\right) = b^k + \left(X'VX\right)^{-1}X'\left(Y - \hat{Y}^k\right)$$

Note : Dans le cas où l'on a une seule variable X et que celle-ci est aussi une variable dichotomique (i.e. 0-1), alors on peut montrer que les coefficients b_0 et b_1 du modèle s'obtiennent directement de :

$$b_0 = \ln\left(\frac{n(x=0,y=1)}{n(x=0,y=0)}\right) \quad b_1 = \ln\left(\frac{n(x=1,y=1)}{n(x=1,y=0)}\right) - b_0$$

où n(x=0,y=1) représente le nombre de fois où un x=0 et un y=1 ont été observés simultanément.

Test d'ajout de variables au modèle

On peut tester l'ajout de variables au modèle en comparant le modèle réduit au modèle complet. Le modèle réduit a "r" paramètres, le modèle complet "c" paramètres. Alors $2(\ln(L_c)-\ln(L_r))$ est distribué suivant une Khi-deux avec "c-r" degrés de liberté.

Intervalle de confiance sur le "odds-ratio"

Un intervalle de confiance sur le "odd-ratio" peut être construit en prenant simplement $[\exp(\hat{W}_{inf}), \exp(\hat{W}_{sup})]$ où $\hat{W}_{inf}$ et $\hat{W}_{sup}$ sont les limites de l'intervalle de confiance sur W (voir section 2.4.2 iii.).

Tableau de classification

Si $\hat{Y}$ >0.5, alors on devrait considérer que l'observation est de type 1, sinon de type 0. Connaissant pour ces observations le type réel obtenu, on peut construire un tableau de contingence 2 x 2 donnant les résultats de la régression logistique.

		Groupe classé	
		0	1
Groupe observé	0	25	10
	1	3	23

Le taux de bonne classification est (25+23)/(25+23+3=10)=78.7%

Des comparaisons entre analyse discriminante (cas de 2 groupes) et régression logistique ont montrées généralement de meilleures classifications par régression logistique.

Résidus

Ici, on considère le cas où au moins une des variables X est continue. Pour le cas où tous les X sont des variables de type "catégorie", des modifications existent (voir Hosmer et Lemeshow, 1989, Applied logistic regression).

Contrairement à la régression linéaire, la variance dépend de la probabilité (W). On utilisera donc des résidus normalisés en fonction de la variance. Deux résidus différents ont été proposés dans la littérature:

Résidu de Pearson:

$$r_j = \frac{Y_j - \hat{Y}_j}{\sqrt{\hat{Y}_j(1-\hat{Y}_j)}}$$

et le résidu déviance :

$$d_j = \sqrt{2\left|ln(\hat{Y}_j)\right|} \text{ si } Y_j = 1$$
$$d_j = -\sqrt{2\left|ln(1-\hat{Y}_j)\right|} \text{ si } Y_j = 0$$

Note: La somme des carrés des résidus déviance est égale à 2ln(L).

Influence des observations

Une mesure analogue à celle utilisée pour la régression linéaire permettant de mesurer l'influence de chaque observation sur la détermination du modèle est fournie par:

$$D_i = (b_{(i)} - b)'(X'VX)(b_{(i)} - b)$$

Les observations ayant une forte influence doivent être examinées avec soin.

Ex. Gonflement des remblais sous les dalles de résidences: PFE de P.A. Pasquier(1999)

La présence de pyrite dans les remblais que l'on rencontre sous les résidences fournit le souffre nécessaire à la formation de gypse. Un gonflement du remblai s'ensuit qui peut occasionner la fracturation de la dalle, le déplacement des murs et donc d'importants dommages aux résidences. P.A. Pasquier (1999) a effectué le relevé de plus de 226 rapports d'expertises réalisés chez LVM-Fondatec. Il a noté plusieurs variables liées au domicile et au remblai (âge de la maison, polyéthylène sous la dalle, calibre du remblai, présence de fissures, déformation des murs, IPPG du remblai (voir CTQ-M-100), épaisseur de la dalle, qualité du béton, ...). Il a codé une variable (Y) prenant la valeur 0 ou 1 selon qu'il y ait présence ou absence de dommages notables. Il s'agit d'un contexte idéal de régression logistique (ou d'analyse discriminante) pour prédire la probabilité qu'un domicile éprouve des dommages.

Ex. Acceptation-rejet de géotextiles

Un géotextile doit passer différents critères de performance pour être accepté comme matériau pour la construction de routes. Pour ce faire, on soumet des géotextiles, possédant des ouvertures de filtration différentes, à des tests de perméabilité où l'on change la nature du sol en contact avec le géotextile (type de sol, profil de la courbe granulométrie, uniformité, stabilité interne du sol), la fréquence et l'importance des cycles de drainage et la charge morte ou surcharge imposée au système, etc. Pour chaque géotextile, on applique la règle de décision pour savoir si celui-ci est acceptable. On cherche ensuite à établir l'effet et l'importance des différentes variables de contrôle dans la décision rendue.

2.9.3 Autres sujets

Transformations

Dans bien des cas, un problème peut être linéarisé par transformation. On applique alors la régression aux variables transformées. Autant que possible, il est préférable d'effectuer les transformations sur X et non sur Y. En transformant Y, la solution obtenue est optimale pour la variable transformée, pas pour la variable Y elle-même. La transformation inverse peut induire des problèmes de biais et de non-optimalité.

Moindres carrés pondérés

Lorsque les variances des résidus ne sont pas égales, on peut donner un poids à chaque observation dans la régression. Ces poids sont habituellement les inverses des variances des résidus. Ceci a pour effet de normaliser les résidus en fonction de leur variance. Notons qu'il suffit parfois de transformer les variables X (et/ou Y) pour stabiliser les variances. La transformation logarithmique est souvent utilisée en ce sens.

Soit une matrice de poids des observations W, les coefficients de la régression seront alors donnés par:

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

Note: Ici W est une matrice diagonale et habituellement $W=V^{-1}$ où V est la matrice diagonale contenant les variances des résidus.

Moindres carrés généralisés

Généralise l'idée précédente dans le cas ou les résidus sont corrélés entre eux, i.e. la pondération tient compte à la fois des variances et des covariances entre résidus. Le problème de la détermination de la matrice de covariance V des résidus est assez complexe et requiert habituellement des procédures itératives. Une fois V connu, les équations sont presque identiques au cas classique.

En posant $W=V^{-1}$, les équations précédentes pour le cas pondéré demeurent valides. La seule différence ici est que W n'est plus nécessairement une matrice diagonale.

Ces deux dernières techniques peuvent être considérées comme la recherche d'une transformation linéaire simultanée sur Y et X qui permet d'obtenir des résidus non-corrélés et de variance égale. Une fois ce résultat obtenu, on applique le moindre carré ordinaire aux variables transformées.

Multicollinéarité

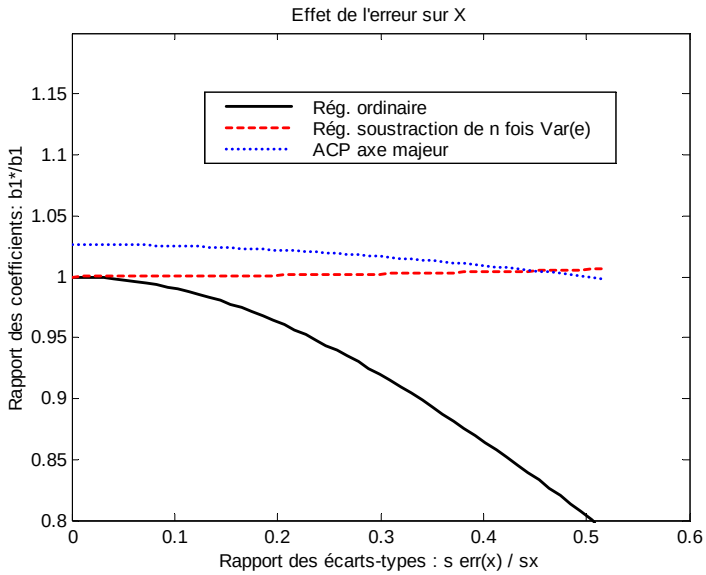
Lorsque les variables X sont très corrélées, il peut arriver que X'X soit quasi-singulière. Dans un tel cas, il faut éliminer une (ou plusieurs) des variables X pour éliminer la multicollinéarité. La conséquence de conserver toutes les variables est des estimés très instables des b. Plusieurs méthodes existent pour détecter des conditions de singularité, habituellement basées sur la détermination des valeurs propres de la matrice X'X. On peut aussi utiliser dans ce cas une régression basée sur les composantes principales (voir chap.3). Une autre stratégie consiste à imposer des contraintes sur les coefficients de la régression (c'est ce qui est fait dans les programmes spécialisés d'analyse de variance que l'on rencontre dans l'étude d'expériences planifiées). Normalement, en utilisant des procédures de sélection avant des variables, l'on ne rencontre pas ce problème car la variable très corrélée aux variables déjà dans la régression ne peut être sélectionnée car l'information qu'elle contient relativement à Y est prise en compte déjà par les autres variables.

Erreur pure et manque d'ajustement

Lorsqu'on a plusieurs observations en un certain nombre de valeurs X_i, on peut séparer la SCE en deux parties, celle due à l'erreur pure (SCE_p) et celle due au manque d'ajustement du modèle (SCE_a). La SCE_p est calculée en prenant la somme des carrés des résidus par rapport à la moyenne des résidus pour chaque X_i dont on dispose de plusieurs observations. Le nombre de degrés de liberté de SCE_p est alors égal au nombre total d'observations avec X répétés - nombre de valeurs différentes où des répétitions sont disponibles. Ainsi, si on a répété 4 fois à X=2.3 et 3 fois à X=4.1 dans une régression comprenant au total 25 observations et 6 variables, on aura (4+3)-2= 5 d.l. pour SCE_p (et 25-5-6-1) =13 d.l. pour SCE_a). On peut tester le manque d'ajustement par rapport à l'erreur pure. Pour plus de détails, voir Draper et Smith).

Variables explicatives sujettes à erreur

Quand les variables explicatives sont aussi sujettes à erreur, les coefficients estimés par régression sont biaisés. Les choses deviennent beaucoup plus compliquées sauf pour le cas où la erreurs sur les X sont de faible amplitude par rapport aux variations des X eux-mêmes. La figure suivante montre que pour des valeurs réalistes de l'écart-type de l'erreur sur X par rapport à l'écart-type de X, l'estimé de b est fiable. En effet, même avec une erreur de 20%, le b estimé demeure à 95% du b vrai.



Coefficient estimé par la régression en fonction de l'importance de l'erreur sur X (1000 observations, vrai modèle : Y=1+3*X+e (e~Normale(0,2500)). Estimation fait par régression, par ACP et par régression en supposant que l'on connaisse la variance de l'erreur sur x.

Lorsque l'erreur sur X n'est pas négligeable, l'on peut utiliser le dernier vecteur propre d'une ACP de la matrice des covariances ou de la matrice des corrélations pour déduire un meilleur estimé de l'équation de régression (voir Chapitre 3). Si l'on connaît la variance des erreurs sur X, on peut soustraire ce terme de la matrice X'X pour calculer un meilleur estimateur « b » de la vraie relation liant Y et X.

Note : Malgré que la régression donne un estimé biaisé du coefficient « b » liant x et y lorsque x est entaché d'erreur, l'équation de prédiction obtenue avec la régression n'en demeure pas moins celle qui est la plus précise. Ainsi, dans l'exemple précédent lorsque s(erreur(x))/s(x)=0.5, la somme des carrés des erreurs de prédiction est 24% plus élevée avec la régression corrigée et 23% plus élevée avec l'ACP qu'avec la droite de régression biaisée. Si l'objectif est de prédire Y (avec un X qui est entaché d'erreur), on utilise la droite de régression biaisée. Si l'objectif est de décrire le lien « véritable » existant entre le X sans erreur et le Y, alors on doit utiliser les coefficients d'une des 2 autres méthodes.

Réponses aux questions et exercices

Question 1

ajout d'une constante: aucun effet
multiplication par une constante c:
variance: multipliée par c^2
covariance: multipliée par c
corrélation: inchangée.

Question 2

Présence d'une donnée extrême qui cause toute la corrélation. Il peut s'agir d'une donnée erronée, mais pas nécessairement. Il faudrait trouver des valeurs de (x,y) intermédiaires.

Question 3

Poser $x_{tr}=(x-10)^2$

Question 4

- si $r_{xy}=0$
  $s_{xy}=0$
  or $b_{y|x}=s_{xy}/s_x^2=0$
  et $b_{x|y}=s_{xy}/s_y^2=0$
  Les deux droites seront donc orthogonales

- si $r_{xy}=1$, les deux droites sont confondues.

- plus la corrélation entre x et y augmente, plus l'angle entre les deux droites diminue. On peut montrer à l'aide d'une simple construction géométrique que l'angle entre les 2 droites vaut :

$$90-\text{atan}(b_{x|y})-\text{atan}(b_{y|x})$$

- non car le critère des moindres carrés définit deux droites différentes. Si la relation $c_1=1/b_1$ tenait, cela voudrait dire que les deux droites sont confondues.

Question 5

Il suffit d'enlever la colonne de "1" dans X.

Question 6

$$R^2 = \frac{SCR_m}{SCT_m} = \frac{SCR_m}{(SCR_m + SCE)}$$

$$\frac{1}{R^2} = 1 + \frac{SCE}{SCR_m}$$

$$\frac{1 - R^2}{R^2} = \frac{SCE}{SCR_m}$$

$$\frac{R^2(n - p - 1)}{(1 - R^2)p} = \frac{SCR_m / p}{SCE / (n - p - 1)} \sim F_{p,(n-p-1)}$$

où p: nombre de variables (p+1 paramètres si on inclut la constante) et n est le nombre d'observations.

Question 7

La régression avec les variables centrées est équivalente à celle pour les variables non-centrées sauf qu'il n'y a pas de constante b_0 dans le modèle centré.

Supposons X et Y centrées. Lorsqu'il n'y a qu'une variable X, on sait que:

$b_1 = s_{xy} / s_x^2$

$$R^2 = \frac{SCR}{SCT} = \frac{Y'Y_p}{Y'Y} = \frac{Y'Xb1}{s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r^2$$

Question 8

Le variogramme. Il suffirait d'ordonner les résidus selon un (ou plusieurs) critère (ex. temps, coordonnées spatiales, méthode de mesure, variable X...). Dans le calcul du variogramme, ce critère joue le même rôle que les coordonnées spatiales habituellement.

Question 9

Il ne peut y avoir plus de paramètres à estimer que de données. On ne peut donc inclure plus de n-1 variables X (plus la constante b_0. Si on en inclut n-1, alors on aura nécessairement $R^2 = 1$. Cela ne veut pas dire que le modèle est bon.

Question 10

On a nécessairement $R^2_{p+1} \geq R^2_p$

Question 11

2^p sous-ensembles différents incluant l'ensemble vide

Question 12

En extrapolation, des polynômes d'ordre élevé peuvent donner des résultats tout à fait farfelus (concentrations négatives, excédant 100%, ...)

Question 13

On définit une variable indicatrice par type de roche. Si on a une constante dans le modèle, alors on définit p-1 variables indicatrices. Le p ^{ième} type de roche s'obtient en posant toutes les variables indicatrices à 0. Si les types de roche ont une séquence logique (ex. basaltes, andésites, rhyolites), on peut aussi parfois les coder par une seule variable quantitative. Ce dernier modèle est un peu moins flexible car il comporte moins de paramètres.

Question 14

Connaissant la correlation partielle, on peut tester le caractère significatif de l'ajout d'une variable. Or, il n'est pas nécessaire d'effectuer la régression explicitement pour ce faire, il suffit de connaître la matrice de corrélation simple. À partir de celle-ci, on peut calculer toutes les corrélations partielles à l'aide d'une formule récursive simple et très rapide à calculer.

Questions 15 et 16

r=-1. Si l'épaisseur relative d'une unité croit, l'autre doit décroître de la même valeur.

Question 17

Corrélations négatives. Plus il y a de variables, moins cet effet est important.

Exercice 1

$$X'X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \quad X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} = \frac{1}{n^2 s_x^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X'X)^{-1} X'Y = \frac{1}{n^2 s_x^2} \begin{bmatrix} (\sum Y_i)(\sum X_i^2) - (\sum X_i)(\sum X_i Y_i) \\ -(\sum X_i)(\sum Y_i) + n(\sum X_i Y_i) \end{bmatrix}$$

or,

$$(\sum Y_i)(\sum X_i^2) - (\sum X_i)(\sum X_i Y_i) = n\overline{Y}(ns_x^2 + n\overline{X}^2) - n\overline{X}(ns_{xy} + n\overline{X}\,\overline{Y}) = n^2\overline{Y}s_x^2 - n^2\overline{X}s_{xy}$$

$$-(\sum X_i)(\sum Y_i) + n(\sum X_i Y_i) = n^2 s_{xy}$$

d' où :

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \overline{Y} - \dfrac{s_{xy}}{s_x^2}\overline{X} \\ \dfrac{s_{xy}}{s_x^2} \end{bmatrix}$$

d'où on tire que

$$b_0 = \overline{Y} - b_1\overline{X}$$

Ce qui indique que la régression passe nécessairement par la moyenne. Ce résultat se généralise d'ailleurs au cas multivariable et on peut toujours obtenir b_0 à partir des autres coefficients $b_1 ... b_p$ par:

$$b_0 = \overline{Y} - [b_1, b_2, ...b_p] \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ . \\ . \\ . \\ \overline{X}_p \end{bmatrix}$$

Le fait que la régression passe par la moyenne est assuré puisque la somme (et donc la moyenne) des résidus donne toujours 0 pour les modèles avec constante (déjà démontré). Comme e=Y-Xb, le résultat découle.

On peut donc toujours effectuer la régression en utilisant les variables centrées et un modèle sans constante. On obtient ainsi les $b_1...b_p$ puis on calcule le b_0 pour utiliser avec les variables non-centrées.

Exercice 2

SCT= Y'IY I est idempotente
SCM= Y'(11'/n) Y où 1 est un vecteur de 1 (n x 1).
(11'/n) (11'/n) = 11'/n
SCT_m= (I-11'/n)

et (I-11'/n) 11'/n = 0

... ainsi de suite pour les autres relations

Exercice 3

e'**1**=0 et e'Y_m=0
e=Y-Xb=Y-X(X'X)$^{-1}$X'Y
e'**1**=Y'**1**-Y'X(X'X)$^{-1}$X'**1**
or
(X'X)$^{-1}$X'**1**= [**1** 0...0]' car le vecteur **1** est la première colonne de X et par la définition d'une inverse.
et
X[1 0...0]'=1
donc
e'1=0

pour e'Y_m=0, la démonstration est identique puisque Y_m=**1**m

e'Y_p=0
Y_p=Xb
e'Y_p=(Y-Xb)'Xb=Y'X(X'X)$^{-1}$X'Y - Y'X(X'X)$^{-1}$X'X(X'X)$^{-1}$X'Y = 0

Y'Y_p=Y_p'Y_p
Y'Xb=Y'X(X'X)$^{-1}$X'Y=Y'X(X'X)$^{-1}$X'X(X'X)$^{-1}$X'Y=Y_p'Y_p

e'X=0
(Y-Xb)'X=Y'X-Y'X(X'X)$^{-1}$X'X=0

Donc, le vecteur de résidus est orthogonal à chaque colonne de la matrice X, i.e. il est dans un espace différent. Comme Y=Y_p+e, il résulte que Y'X=Y_p'X dont les résultats ci-haut ne représentent que quelques cas particuliers.

Exercice 4:

Considérons le cas général où on veut tester
H_0 moyenne de Y=m
vs H_1 moyenne de Y≠m
On pose Y_c=Y-m Y_c est alors de moyenne 0 sous H_0
On utilise le modèle Y_c=b_0+e
On effectue la régression et on teste H_0: b_0=0

On calcule: F=(SCR/1) / (SCE/(n-1))
La statistique F est alors comparée à une $F_{1,(n-1)}$
Ce test est identique au test de student vu au cours 327.

Exercice 5

Au lieu d'effectuer le test avec SCR_m, on utilise SCR.

Exercice 6

On place dans l'ordre les variables x, y et z. On calcule par la suite:

$$\Sigma_{11\bullet 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{bmatrix} s_x^2 - \frac{s_{xz}^2}{s_z^2} & s_{xy} - \frac{s_{xz}\,s_{yz}}{s_z^2} \\ s_{xy} - \frac{s_{xz}\,s_{yz}}{s_z^2} & s_z^2 - \frac{s_{yz}^2}{s_z^2} \end{bmatrix}$$

La corrélation partielle s'obtient en prenant la covariance partielle et en divisant par les variances partielles. Après quelques simplifications, on retrouve l'expression 2.27

Exercice 7

On compare:

$$(n-p-1)\,r^2_{yx_p|x_1,x_2,\ldots,x_{p-1}}$$

à une $F_{1,n-p-1}$

Note: Lorsque p=1, on retrouve le test classique pour les coefficients de corrélation simple vu au cours MTH2301.