8. Régression

8.1 *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

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad 8.6$$

où Y est la variable à expliquer,
X est la variable explicative,
ε est le résidu théorique entre la droite et Y

Dans cette équation, β_0 et β_1 représentent les paramètres (vrais) de la droite de régression.

On estime les coefficients β_0 et β_1 inconnus par la méthode des moindres carrés. On peut montrer que les coefficients b_0 et b_1 sont donnés (dans le cas de la régression de y sur x) par (voir section 8.2):

$$\mathbf{b}_0 = \overline{\mathbf{y}} - \mathbf{b}_1 \overline{\mathbf{x}}$$
$$\mathbf{b}_1 = \frac{\mathbf{s}_{xy}}{\mathbf{s}_x^2} \qquad 8.7$$

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

Remarque: A proprement parler, la droite précédente est la droite des moindres carrés et non la 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 (i.e. X n'est pas une v.a.). Il suffit alors que les résidus du modèle (les ε) 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.

8.2 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.

Qu'entend-t-on par modèle linéaire ?

On entend par modèle linéaire tout modèle dont les coefficients « b » peuvent être obtenus par solution d'un système d'équations linéaires, soit directement, soit après transformation.

Exemples de modèles linéaires :

$Y=b_0+b_1X+e$
$Y=b_0+b_1X_1+b_2X_2+b_3X_3+e$
$Y= b_0+b_1X+b_2X^2+b_3X^3+e$
$Y=b_0+b_1\log(X)+e$
$Y=b_0+X^2\exp(b_1)+e$
$Y= b_0X_1^{b_1}X_2^{b_2}e$ (en prenant le log : $\ln(Y)=\ln(b_0)+b_1\ln(X_1)+b_2\ln(X_2)+\ln(e)$)
$Y=b_0+b_1X_1+b_1^{b2}X_2+e$

Exemples de modèles non-linéaires :

$Y= b_0 + X_1^{b_1}X_2^{b_2} + e$
$Y=b_0X^{b1}+e$
$Y=b_0+X^{b1}+e$

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

$$Y = \beta_0+\beta_1 X_1+\beta_2 X_2+ .....+\beta_p X_p+\varepsilon \qquad 8.12$$

On cherche à estimer les p+1 coefficients β_0, β_1,...., β_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. On écrit alors:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & X_{1p} \\ 1 & X_{21} & X_{22} & . & . & X_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & X_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} \qquad 8.13$$

Ou, plus simplement:

$$Y = Xb + e \qquad 8.14$$

La somme des carrés des erreurs s'écrit:

$$SCE = e'e = (Y - Xb)'(Y - Xb) \qquad 8.15$$

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:

$$SCE = Y'Y - Y'Xb - b'X'Y + b'X'Xb \qquad 8.16$$

$$\frac{\delta SCE}{\delta b} = 0 = (X'X)b - X'Y \qquad 8.17$$

d'où on tire finalement:

$$b = (X'X)^{-1} X'Y \qquad 8.18$$

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.

8.2.1 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
$Y'Y_p=Y_p'Y_p$

8.2.2 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 une variable que l'on peut contrôler, 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. la régression explique quelque chose (en sus de la moyenne).

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 entre 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:

$$Y_p \pm t_{n-(p+1),\alpha/2}\ s(1+\mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})^{0.5} \qquad \text{avec } s=[SCE/(n-(p+1))]^{0.5}$$

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:

$$Y_p \pm t_{n-(p+1),\alpha/2}\ s(1/k + \mathbf{x_i}(X'X)^{-1}\mathbf{x_i'})^{0.5}$$

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:

$$Y_p \pm t_{n-(p+1),\alpha/2}\ s(\mathbf{x_i}(X'X)^{-1}\mathbf{x_i}')^{0.5}$$

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: $b_i \pm t_{n-(p+1),\alpha/2}\ s_{bi}$ 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 sur la diagonale.

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

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 8.2.7

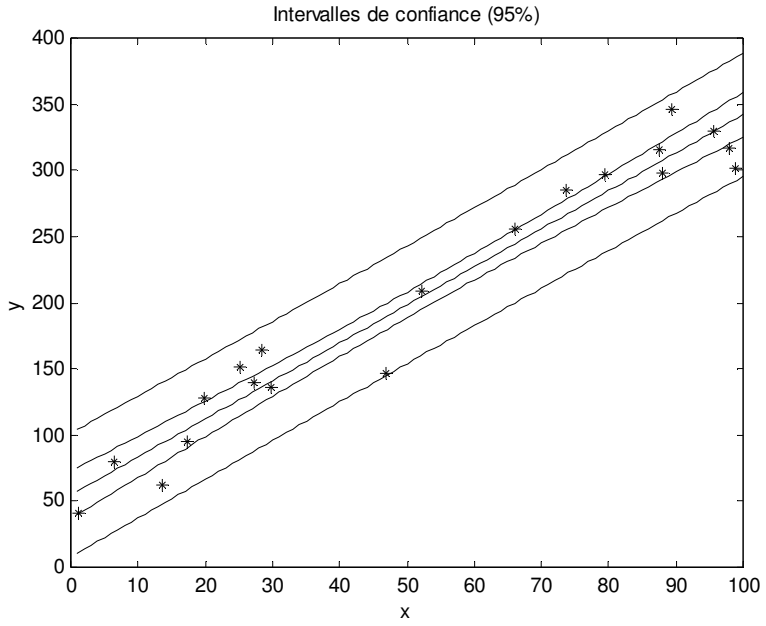
vi. Vous voulez vérifier si deux ou plusieurs coefficients de la régression sont égaux.

Voir section 8.2.7

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

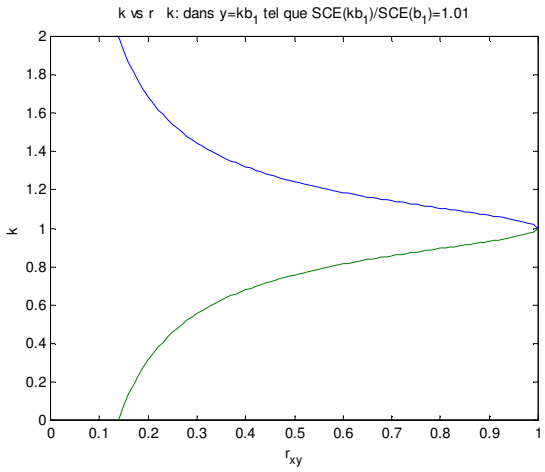
Voir section 8.2.7

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.



8.2.2.1 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.



8.2.3 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:

$$R^2 = \frac{SCR_m}{SCT_m} \qquad 8.19$$

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:

$$R^2 = 1-\frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2} = 1-\frac{SCE}{SCT_m} \qquad 8.19b$$

Note: Les deux expressions précédentes sont équivalentes dans le cas d'un modèle avec constante.
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 importante concernant les transformations : Lorsque l'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 8.19b.

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.

8.2.3.1 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	$Y_p \pm t_{n-2,\alpha/2}\ s\left\{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$
Intervalles de prédiction $(Y\|x_0)$	$Y_p \pm t_{n-(p+1),\alpha}\ s(1+\mathbf{x_0}(X'X)^{-1}\mathbf{x_0}')^{0.5}$	$Y_p \pm t_{n-2,\alpha/2}\ s\left\{1+\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$
Intervalle de prédiction pour la moyenne de k observations	$Y_p \pm t_{n-(p+1),\alpha}\ s(1/k+\mathbf{x_0}(X'X)^{-1}\mathbf{x_0}')^{0.5}$	$Y_p \pm t_{n-2,\alpha/2}\ s\left\{\frac{1}{k}+\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$
Intervalle de confiance sur β_0	n.a.	$b_0 \pm t_{n-2,\alpha/2}\ s\left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$
Intervalle de confiance sur β_1	n.a.	$b_1 \pm t_{n-2,\alpha/2}\ s\left\{\frac{1}{\sum(x_i-\bar{x})^2}\right\}^{1/2}$

Note : σ^2 est estimé en pratique par $CME=SCE/(n-p-1)$, $s=CME^{1/2}$; t : Student.

8.2.4 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...)

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. L**'influence** des données peut aussi être utilisée (voir section 8.2.4.2)

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 $Var(e) = \sigma^2(I-M)$, 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").

8.2.4.1 Transformations des données

Transformation sur Y
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.

Transformation d'un (ou plusieurs) X

Ces transformations ne causent aucun problème. On considère souvent des transformations trigonométriques, logarithmiques, un polynôme en x, etc.

8.2.4.2 La notion d'influence d'une observation

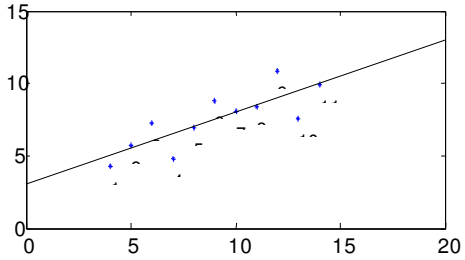
Lorsqu'on effectue une régression, il est important de vérifier si le modèle obtenu peut être causé par une (ou quelques unes) observation particulière. On espère habituellement que le modèle représente une caractéristique générale des données et non l'influence d'une seule donnée particulière. L'examen 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.

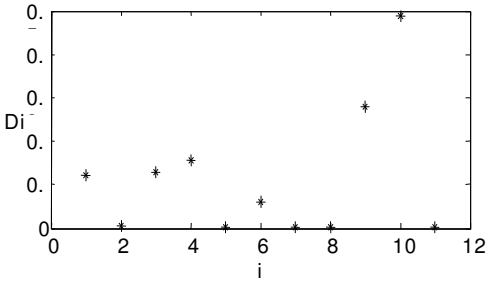
La figure suivante montre 4 ensembles de données ayant les mêmes coefficients "b", les mêmes R^2 et les mêmes CME. Pourtant, seul le 1er modèle est adéquat. On peut mesurer l'influence d'une observation à l'aide de la distance suivante (distance de Cook):

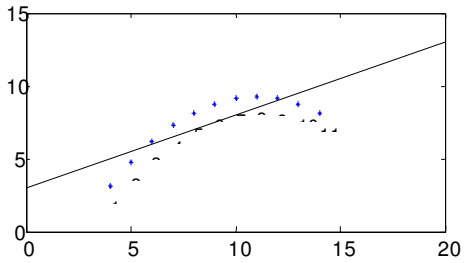
$$D_i = \frac{(b_{(i)} - b)'(X'X)(b_{(i)} - b)}{(p+1)CME} = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{(p+1)CME} \qquad 8.20$$

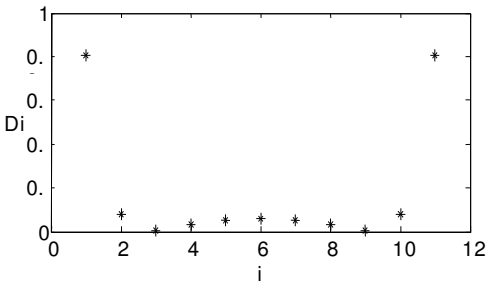
La notation (i) signifie que la $i^{ème}$ observation est enlevée. Weisberg (1985) indique que les observations présentant un D_i supérieur à 1 sont très influentes et doivent être examinées avec attention.

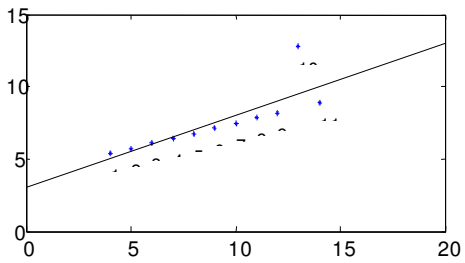
La figure montre, à droite, les D_i associées à chaque observation des données de gauche. Les 4 régressions ont exactement les mêmes coefficients b et le même $R^{2.}$

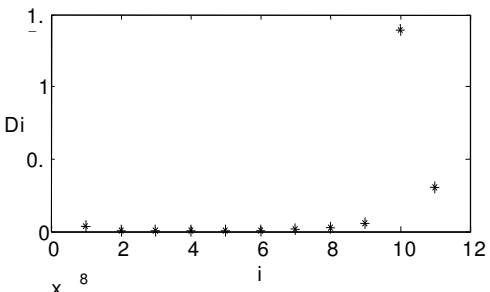

15
10
5
0
0
5
10
15
20

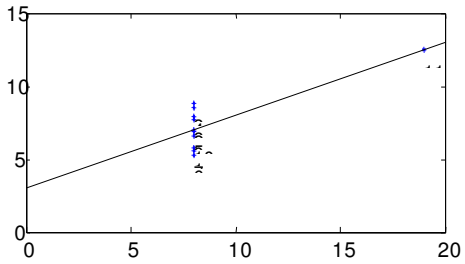

Di
0
2
4
6
8
10
12
i

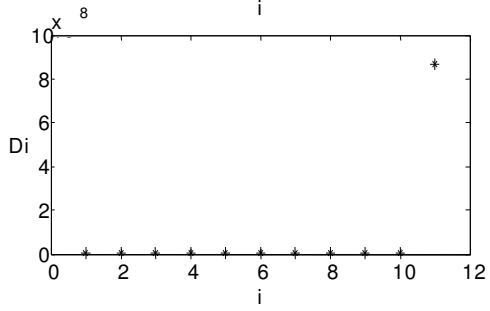

15
10
5
0
0
5
10
15
20


Di
0
2
4
6
8
10
12
i


15
10
5
0
0
5
10
15
20


Di
0
2
4
6
8
10
12
i


15
10
5
0
0
5
10
15
20


Di
0
2
4
6
8
10
12
i

8.2.5 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):

$$Y = X_r \beta_r + \varepsilon_r \qquad 8.21$$

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

$$Y = X_c \beta_c + \varepsilon_c \qquad 8.22$$

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):

$$\frac{(SCE_r - SCE_c)/k)}{SCE_c/(n-p-1)} \sim F_{k,(n-p-1)} \qquad 8.23$$

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.

8.2.5.1 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 presque 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 $(1-.05/50)^{50}\approx 0.95$, $(1-.05/30)^{30}\approx 0.95$.

8.2.5.2 "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:

$$\%\,Na_2O = \beta_{00} + \beta_{10}x + \beta_{01}y + \beta_{20}x^2 + \beta_{11}xy + \beta_{02}y^2 + \ldots + \beta_{0k}y^k + \varepsilon \qquad 8.25$$

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 coïncident 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 définit 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. 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.

8.2.5.3 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 possibles 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.

8.2.6 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.

8.2.6.1 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 jours précédant le relevé.

TM: température moyenne du mois où 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

$$\frac{(SCE_r - SCE_c)/2}{SCE_c/(n-3)} \qquad 8.26$$

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.

8.2.7 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.

8.2.7.1 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]

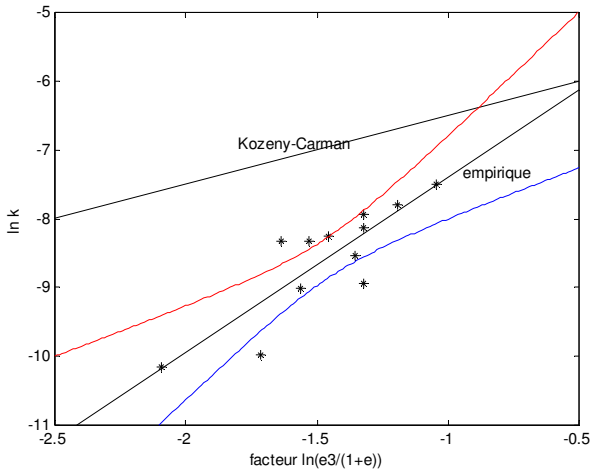
On a le modèle: $\ln(k)=b_0+b_1*\ln(e^3/(1+e))$

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.



8.2.7.2 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 $t_{23-1;.95}$=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.

8.2.7.3 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
$u=r^2S / 4Tt$
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, 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

Ces valeurs sont très semblables à celles obtenues par régression.

8.2.7.4 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:

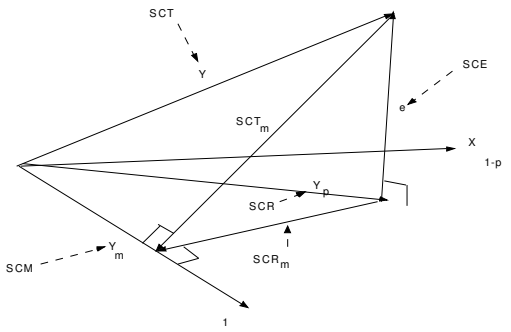
$Q_{j+1} = r_a Q_j$ j après la période de crue
$Q_{j+1} = r_t Q_j$ j en période de crue avant la transition

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.

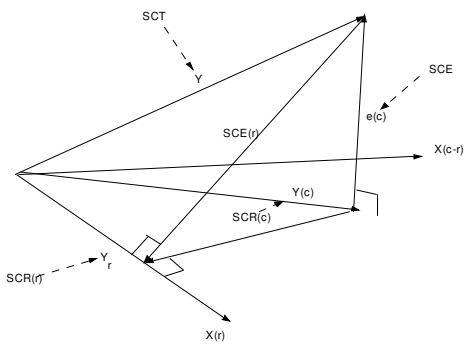
8.3 Géométrie des moindres carrés

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)



8.4 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.

8.5 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.

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. Plusieurs méthodes existent pour détecter des conditions de singularité, habituellement basées sur la détermination des valeurs propres de la matrice X'X. La conséquence de conserver toutes les variables est des estimés très instables des b (la variance de b devient très grande et des covariances négatives très grandes entre certains coefficients apparaissent).

Dans un tel cas, on peut régler le problème de diverses façons :

- en utilisant des procédures de sélection avant des variables, où l'on ne rencontre pas ce problème car la variable très corrélée aux variables déjà dans la régression ne peut normalement être sélectionnée car l'information qu'elle contient relativement à Y est déjà prise en compte par les autres variables ;
- en retirant une (ou plusieurs) variables X ;
- en imposant 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).
- en transformant les X pour qu'ils deviennent orthogonaux (ACP : analyse en composantes principales) et en ne retenant qu'un sous-ensemble des « p » nouvelles variables ;
- en recourant à une régression biaisée (« ridge regression »). Ceci est obtenu en ajoutant une perturbation positive sur la diagonale de la matrice X'X.

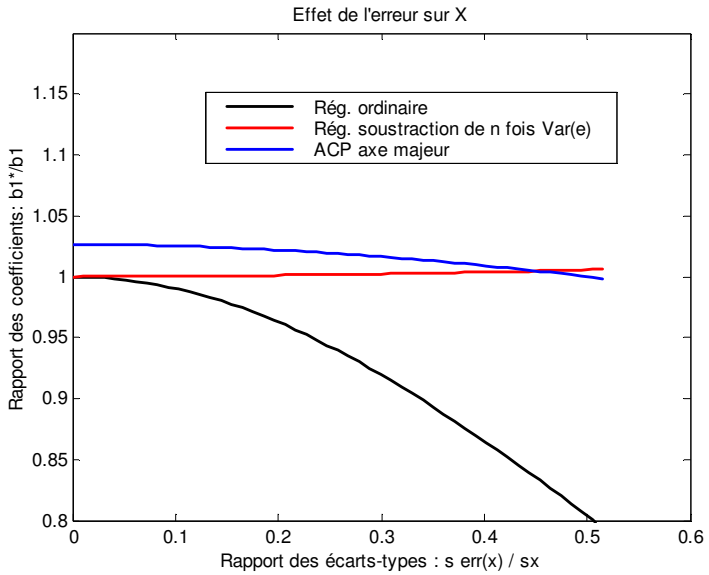
Une statistique utilisée pour détecter la multicollinéarité est le facteur d'inflation de la variance (Variance Inflation Factor en anglais). On le définit comme $1/(1-R_j^2)$ où R_j^2 est le coefficient de détermination obtenu en effectuant la régression de la variable X_j sur les autres variables explicatives X. On considère qu'un VIF supérieur à 10 indique un problème possible de multicollinéarité.

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.

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.

8.6 Réponses aux questions et exercices

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 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.

Exercice 1

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

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

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

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

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

$$b0 = \overline{Y} - b1\overline{X}$$

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= Y'(I-11'/n)Y

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.

Exercice 5

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