GLQ3402 -- Examen de mi-session

Jeudi le 16 juin 2005
8h30 à 11h00

Toute documentation permise.
Calculatrice permise. Ordinateur portable interdit.

L'examen comporte 6 questions totalisant 50 points. <u>Chaque réponse doit être accompagnée des calculs et justifications appropriés.</u>

Les questions valent dans l'ordre : 4, 12, 12, 9, 9, 4 points.

Points

4 1- Indiquez pour les modèles suivants s'ils sont i. linéaire, ii. non-linéaire mais pouvant être rendu linéaire par transformation, ou iii. non-linéaire.

a) $Y = X_1^{\beta_1} X_2^{\beta_2} \varepsilon$

b) $Y = \beta_0 + \beta_1 X_1 X_2^{3.1} + \beta_2 X_2 + \varepsilon$

c) $Y = \beta_0 \beta_1^{X_1} \varepsilon$

d) $Y = \beta_1 cos(\beta_2 X_1) + \varepsilon$

---

12 2- Lors de tirs explosifs sur un chantier de construction, un séismographe enregistre la vitesse de déplacement vertical des particules (v en m/s). Vous connaissez la distance entre le séismographe et le tir (d en m) et la charge explosive utilisée (w en kg). Vous effectuez 20 mesures et construisez un modèle de la forme suivante pour faire les prédictions de ln(v) :

$$ln(v) = \beta_0 + \beta_1 ln\left(\frac{w^{1/2}}{d}\right) + e \quad \text{(modèle A)}$$

a) On obtient un R^2 de 0.9, SCE vaut 110. Le coefficient β_1 est-il significatif ? Faites le test requis pour vous en assurer.

Vous effectuez une régression cette fois avec le modèle suivant :

$$ln(v) = \beta_0 + \beta_1 ln\left(\frac{w^{1/2}}{d}\right) + \beta_2 ln(d) + e \quad \text{(modèle B)}$$

b) Le R^2 passe à 0.93. Faites le test permettant de juger si le modèle B est significativement meilleur que le modèle A.

Le client du projet de construction exige de l'exécutant des travaux que les explosions génèrent une vitesse de déplacement vertical inférieure à 100m/s pour un point situé à 20m de distance du point de sautage (ceci pour éviter des endommagements aux structures existantes et au roc de fondation).

c) Expliquez comment vous pourriez utiliser le modèle B pour indiquer à l'exécutant la quantité maximale de charge explosive (w_{max}) pouvant être utilisée pour respecter la norme fixée par le client dans la grande majorité des cas.
(Aide : vous devez accepter un certain niveau de risque de dépasser le seuil fixé; ce risque porte sur la vitesse de chaque tir individuellement et non sur la valeur moyenne de la vitesse pour une charge donnée).

Utilisant le même jeu de données, on calcule cette fois les paramètres du modèle de régression :

$$ln(w) = \beta_0 + \beta_1 ln(d) + \beta_2 ln(v) + e \qquad \text{(modèle C)}$$

d) Votre employeur vous demande en cours de construction de mesurer les vitesses observées, la distance au tir et d'en déduire la charge explosive utilisée pour s'assurer que le constructeur respecte la charge maximale déterminée en c) (Note : l'exécutant peut être tenté d'accroître la charge pour accélérer les travaux). *Quel modèle utiliserez-vous pour faire ce travail, le modèle B ou le modèle C? Pourquoi?*

---

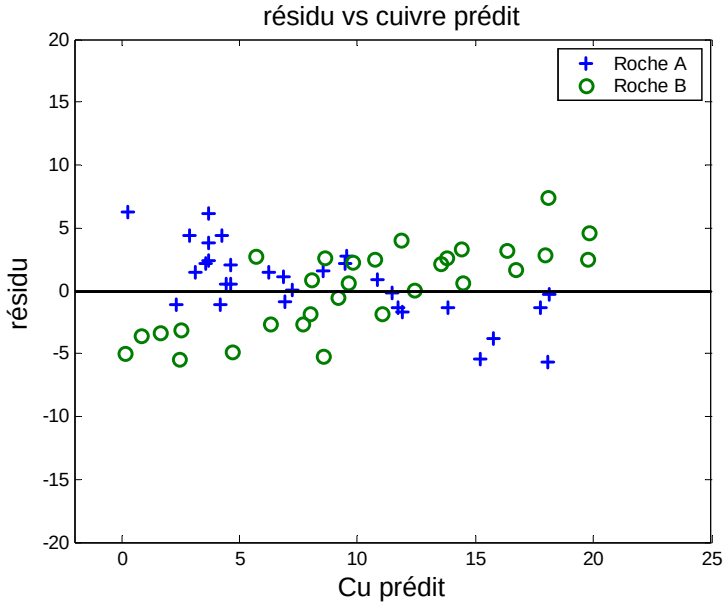
12 3- On veut prédire la teneur en Cu d'une roche à partir des mesures de la conductivité électrique et de la densité obtenues par sondes géophysiques. Le modèle actuel s'écrit :

$$Cu = \beta_0 + \beta_1 \rho + \beta_2 d + \varepsilon$$

où Cu est le % de Cu à l'analyse géochimique, ρ est la résistivité apparente (en ohm-m) et *d* la densité (sans unité). Le Cu se trouve dans des sulfures qui ont la caractéristique d'être plus denses et plus conducteurs que la roche encaissante.

a) Quel est le signe attendu des coefficients b_1 et b_2 ? Quelles sont les unités des coefficients b_1 et b_2 ?

Deux types différents de roche (A et B) sont utilisés dans cette expérience. Le graphe des résidus en fonction des valeurs prédites est construit en utilisant un marqueur différent pour chaque type de roche. On obtient :



b) Indiquez ce qui cloche au niveau des résidus. Suggérez un modèle alternatif au modèle actuel qui pourrait corriger le problème observé (utilisant la même information que disponible présentement et traitant toutes les données en un seul bloc) .

Dans le but de pouvoir distinguer dans les sulfures la portion pyrite de celle chalcopyrite, on a effectué l'analyse géochimique du Fe en plus du Cu. On a obtenu la matrice de corrélation simple suivante :

	Cu	Fe	ρ	d
Cu	1.000	0.285	-0.873	0.734
Fe	0.285	1.000	-0.684	0.786
ρ	-0.873	-0.684	1.000	-0.906
d	0.734	0.786	-0.906	1.000

c) Quel serait le R^2 d'un modèle prédisant le Cu uniquement avec la résistivité ?

d) Calculez la corrélation partielle entre Cu et Fe lorsqu'on fixe la résistivité de la roche. Déduisez ce que serait le R^2 d'un modèle de régression de Cu expliqué par Fe et résistivité (Aide : voir p. 37 équation 2.30). Voyez-vous une limitation pratique à l'utilisation de ce modèle pour prédire le Cu?

---

9 4- On veut effectuer la correction géométrique d'une image satellite déformée. Soit (u,v) les coordonnées sur l'image déformée et (s,t) les coordonnées sur l'image de référence. On réussit à identifier 10 points de contrôle sur l'image de référence et sur l'image déformée. Un modèle polynomial en (s,t) est utilisé pour corriger géométriquement l'image. Pour la suite, on ne s'intéresse qu'à la régression portant sur « u ».

a) La régression $u^*=b_0+b_1s+b_2t$ fournit un SCE de 10. Lorsqu'on considère le modèle quadratique, $u^*=b_0+b_1s+b_2t+b_3s^2+b_4s*t+b_5t^2$, on obtient SCE=8. L'inclusion de la composante quadratique améliore-t-elle significativement le modèle ? Faites le test pour vous en assurer.

b) Quel serait le degré maximum du polynôme que l'on pourrait considérer utiliser dans ce problème en supposant que tous les coefficients du polynôme sont présents dans la régression ? Pourquoi serait-il risqué de retenir un polynôme de degré aussi élevé ?

c) On calcule avec le modèle linéaire de la question a) $\sum_{i=1}^{10} u_i u_i^* = 100$ où u_i^* est la valeur prédite pour le $i^{ième}$ point de contrôle. Que vaut SCT ?

---

9 5- Soit la matrice de covariance partielle suivante entre Y et trois variables X_1, X_2 et X_3. L'effet de deux variables (X_4 et X_5) déjà dans la régression a été éliminé (fixé).

	Y	X_1	X_2	X_3
Y	39	4	2	-10
X_1	4	9	0	2
X_2	2	0	2	0
X_3	-10	2	0	4

a) Dans une procédure de sélection avant, quelle variable sera sélectionnée à la prochaine itération?

b) Quel sera le coefficient de régression associé à cette variable ?

c) Le coefficient calculé en b) demeure-t-il inchangé si l'on poursuit l'algorithme et que l'on inclut une autre variable?

---

4 6- Vous inspirant de l'exemple vu en classe portant sur les intervalles de confiance simultanés (ellipse ou ellipsoïde de confiance sur l'ensemble des coefficients), et les intervalles de chaque coefficient considéré individuellement,

a) est-il possible que tous les coefficients de la régression soient individuellement significatifs mais que la régression dans son ensemble ne soit pas significative ?

b) est-il possible que tous les coefficients, individuellement soient non-significatifs mais que la régression dans son ensemble soit significative ?

Bon examen

Corrigé

1- a) non-linéaire mais peut être linéarisé par transformation logarithmique (si X positif partout)
b) linéaire
c) non-linéaire mais peut être linéarisé par transformation logarithmique.
d) non-linéaire

2- a) R^2=0.9=1-SCE/SCT_m. donc SCT_m=SCE/(1-R^2)=110/0.1=1100. SCR_m=SCT_m-SCE=1100-110=990.
H_0 $\beta_1 = 0$ => test : (SCR_m/1) / (SCE/(20-2))=990/(110/18)= 162. La valeur $F_{table,\ 1,18,.05}$=4.41. Le coefficient est fortement significatif.
b) R^2=0.93=1-SCE/SCT_m; SCE=(1-R^2)SCT_m=0.07*1100=77.
Le test d'ajout est ((110-77)/1) / (77/(20-3))= 7.28 alors que $F_{tabl,1}$, 17,.05=4.45. Le modèle B améliore significativement la prédiction.

c) Ayant les coefficients de la régression, on peut fixer la distance à 20m, et tracer un graphe donnant simplement ln(v) en fonction de ln(w). On peut tracer la limite supérieure de l'intervalle de confiance de niveau 1-α=95% (ou 99% ou une autre valeur raisonnable) pour une valeur observée (p. 11)

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

On trace une ligne horizontale sur ce graphe à ln(100)=4.61. Le point d'intersection avec la limite supérieure indique la charge maximale que l'on peut utiliser tout en encourant un risque $\alpha/2$ de néanmoins dépasser la norme demandée.

d) Le modèle C. Clairement l'objectif ici est de prédire ln(w). C'est donc cette variable qui doit être le Y de la régression. On ne peut utiliser l'autre régression pour faire des prédictions. Comme en c), pour éviter de produire de fausses alarmes, on pourrait construire la borne supérieure de l'intervalle de confiance autour de chaque valeur prédite de ln(w) pour d=20. Si l'intervalle inclut le w_{max} calculé en c) il est possible que client ait utilisé en réalité une charge supérieure à w_{max} (une alarme). Si w_{max} est > la borne supérieure, alors il n'y a pas lieu de suspecter que w_{max} a été dépassé. Sur un grand nombre de tirs « n », on doit s'attendre à ce que $\alpha/2$ *n tirs génèrent des alarmes. Si le nombre observé est significativement supérieur à ce nombre, l'exécutant ne respecte probablement pas la norme fixée.

3- a) b_1 devrait être négatif, b_2 positif. b_1 est en %(ohm-m)$^{-1}$; b_2 est en %.
b) On observe une tendance linéaire dans chaque sous-groupe. Ceci indique que la régression ne présente probablement pas les mêmes coefficients pour les 2 types de roche. Un modèle permettant de tenir compte de cela est :

$$Cu = b_0 + b_1\rho_a + b_2 d + b_3 I_A + b_4\rho_a I_A + b_5 d I_A + e$$

où I_A est une indicatrice prenant la valeur 1 si la roche est de type a et 0 si elle est de type B.

c) 0.873^2=0.762

d) $\dfrac{-0.285-(-0.873)*(-0.684)}{\left((1-0.873^2)*(1-0.684^2)\right)^{0.5}} = -0.877$, R^2=0.762+(1-0.762)*(-0.877)^2=0.945

4 a) test : F_{ajot}= ((10-8)/3) / (8/(10-6))=0.33 comparé à une $F_{table,\ 3,4,0.05}$=6.59. Clairement on ne peut rejeter H_0 donc il ne vaut pas la peine d'introduire la composante quadratique.

b) Comme il y a 10 données, il ne peut y avoir plus de 10 paramètres. On est donc limité à un polynôme d'ordre 3 qui compte exactement 10 termes. Il est dangereux d'utilser ce modèle car alors il ne reste plus aucun degré de liberté pour SCE. On ne peut donc pas juger si ce modèle est significatif. En fait on aura R^2=1 mais on aurait eu le même résultat avec tout ensemble de 10 variables, mêmes choisies aléatoirement.

c) $\sum_{i=1}^{9} u_i u_i^* = 100$ =SCR. Comme SCE vaut 10, SCT vaut 110.

5- on calcule les corrélations (partielles) avec Y.
$r_{Y,X1}=4/(39*9)^{0.5}=0.21$
$r_{Y,X2}=2/(39*2)^{0.5}=0.23$
$r_{Y,X3}=-10/(39*4)^{0.5}=-0.80$

C'est la variable X_3 qui va entrer dans la régression.

b) $b_3= -10/4$

c) Non, tous les coefficients changent lorsqu'on ajoute une variable.

6 a) oui. Avec 2 coefficients par exemple l'ellipse peut inclure la valeur (0,0) et celle-ci peut être hors de l'intervalle individuel sur chaque coefficient
b) oui. La valeur (0,0) peut être en dehors de l'ellipse de confiance et être à l'intérieur de chacun des intervalles individuels.

GLQ3402 -- Examen de mi-session

Jeudi le 10 juin 2004
8h30 à 11h00

Toute documentation permise.
Calculatrice permise

L'examen comporte 4 questions totalisant 50 points. Chaque réponse doit être accompagnée des calculs et justifications appropriés.

Les questions valent dans l'ordre : 17, 18, 9, et 6 points.

Points

1- Vous travaillez comme stagiaire chez Hydro-Québec (HQ) sur un chantier où vous êtes chargés d'effectuer le contrôle de vibrations dues à des tirs de sautage effectués par un entrepreneur indépendant. Votre séismographe enregistre la vitesse de déplacement des particules (v). Vous connaissez la distance par rapport au tir (d) et la charge exacte utilisée pour le tir (w). HQ utilise un modèle de la forme suivante pour faire les prédictions des vitesses :

$$\ln(v) = b_0 + b_1 \ln\left(\frac{d}{w^{1/2}}\right) + e \quad \text{(premier modèle)}$$

1	*a) Quel est le signe attendu du coefficient « b_1 » ?*
2	*b) Décrivez le vecteur Y et les colonnes de la matrice X correspondant à ce problème de régression.*

Après un certain nombre de relevés, vous vous demandez si le modèle utilisé par HQ est le meilleur modèle possible. Vous considérez le modèle alternatif suivant :

$$\ln(v) = b_0 + b_1 \ln(d) + b_2 \ln(w) + e \quad \text{(second modèle)}$$

1	*c) Décrivez le vecteur Y et les colonnes de la matrice X correspondant à ce problème de régression.*
4	*d) Avec les mêmes 30 observations, vous obtenez R^2 =0.8 avec le premier modèle (SCE=2) et R^2=0.9 avec le second modèle. Le second modèle est-il significativement meilleur que le premier ? Posez l'hypothèse H_0 à vérifier et faites le test requis.*
4	*e) L'entrepreneur effectue ses propres mesures de vibrations. Il adopte un modèle identique à votre second modèle (i.e. il adopte un modèle de la forme* $\ln(v) = b_0 + b_1 \ln(d) + b_2 \ln(w) + e$*) mais trouve ses propres coefficients « b ». Expliquez comment vous pourriez tester si ces coefficients sont significativement différents des vôtres. Donnez tous les détails requis (vecteur Y, matrice X, sommes des carrés à utiliser, degrés de liberté du test.)*

Dans les opérations quotidiennes, vous ne connaissez pas la charge « w » réellement utilisée par l'entrepreneur. Le rôle du contrôle des vibrations est justement d'assurer que l'entrepreneur n'utilisera pas des charges excessives (afin d'accélérer les travaux), à l'insu de HQ, risquant ainsi d'endommager le roc (une règle interne de HQ indique que la vitesse des particules à 20m ne doit pas excéder 5cm/s. Tout tir excédant cette règle doit être rapporté à HQ). Vous relevez la vitesse des particules en positionnant le

géophone à une distance variant de 30m à 100m du lieu de tir selon la disponibilité des sites et la sécurité de l'opération. Ayant la vitesse des particules et le modèle précédent (second modèle), vous déterminez la charge « ln(w) » utilisée lors du tir. Ayant cette charge et le modèle, vous calculez la vitesse des particules (ln(v)) que vous auriez enregistré si votre géophone avait été positionné à 20m du tir.

3	*f) Identifiez un problème qui se pose avec cette approche.*

2- On a mesuré la déformation de 17 éprouvettes de béton après une cure humide d'un an. On a aussi déterminé le % de quartz micro-cristallin (chert), le % de ciment siliceux et le % de grains de quartz grossiers pour chaque éprouvette. Ces déterminations ont été réalisées par comptage microscopique. L'objectif de la régression est de prédire la déformation après un an en utilisant les % de quartz comme variables prédictives. La matrice suivante donne les corrélations simples entre les différentes variables.

	Qtz grossier	Qtz fin	Ciment	Déformation
Qtz grossier	1.00	-0.06	0.59	0.28
Qtz fin	-0.06	1.00	-0.26	0.22
Ciment	0.59	-0.26	1.00	0.78
Déformation	0.28	0.22	0.78	1.00

6	*a) Dans une procédure d'inclusion avant, quelles seraient les 2 premières variables à être choisies ? Justifiez en faisant les calculs requis.*
3	*b) Quel serait le R^2 du modèle avec 2 variables ?*

Utilisant le modèle avec les 2 variables sélectionnées, on trouve $Y'\hat{Y} = 0.063$. La moyenne des Y vaut 0.047.

6	*c) Utilisant ces informations et le R^2 déterminé en b) construisez les sommes de carrés : SCT, SCM, SCT_m, SCR, SCR_m, SCE et les degrés de liberté associés. Si vous n'avez pu déterminer le R^2 en b), utilisez la valeur R^2=0.75 (Note : cette valeur n'est pas la réponse pour b), elle est utilisée simplement pour que vous puissiez appliquer la démarche et effectuer les calculs).*
3	*d) L'ajout de la 2^e variable est-il significatif ? Faites le test requis. (Aide : quel serait le R^2 avec seulement la 1ère variable ?)*

3- On dispose de 201 analyses géochimiques des éléments majeurs de roches volcaniques dont 107 sont présumées avoir été altérées (Y=1) par le processus de minéralisation ayant mené à la formation de gisements volcanogènes de sulfures massifs situés à proximité et 94 sont éloignés de tout gisement (Y=0). On applique une régression logistique à ces données et l'on obtient le modèle suivant (après sélection des variables) :

$$\hat{W} = 5.68 - 0.25 * Al2O3 - 0.38 * CaO - 0.41 * Na2O$$

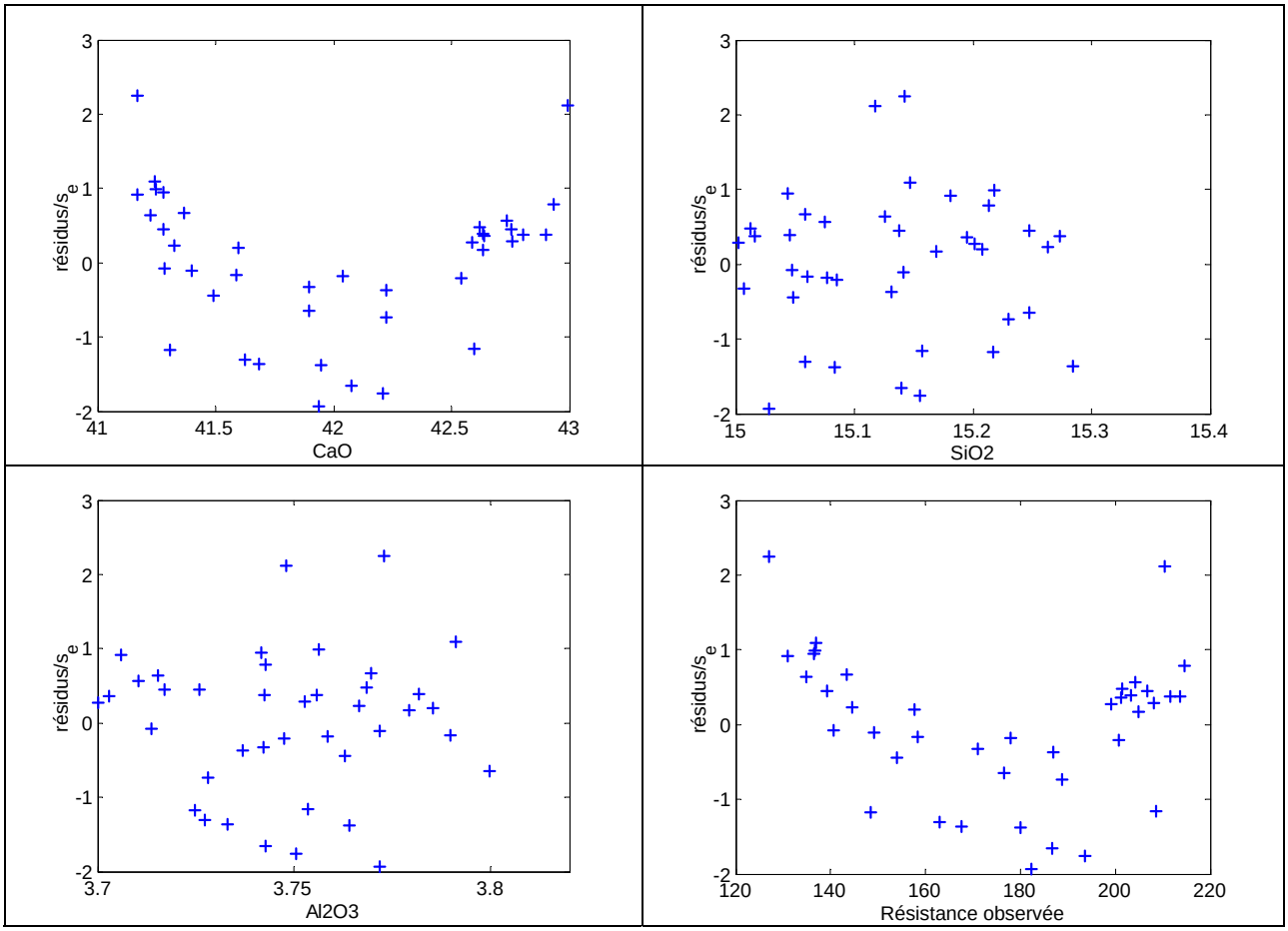
Le tableau de classement est le suivant :

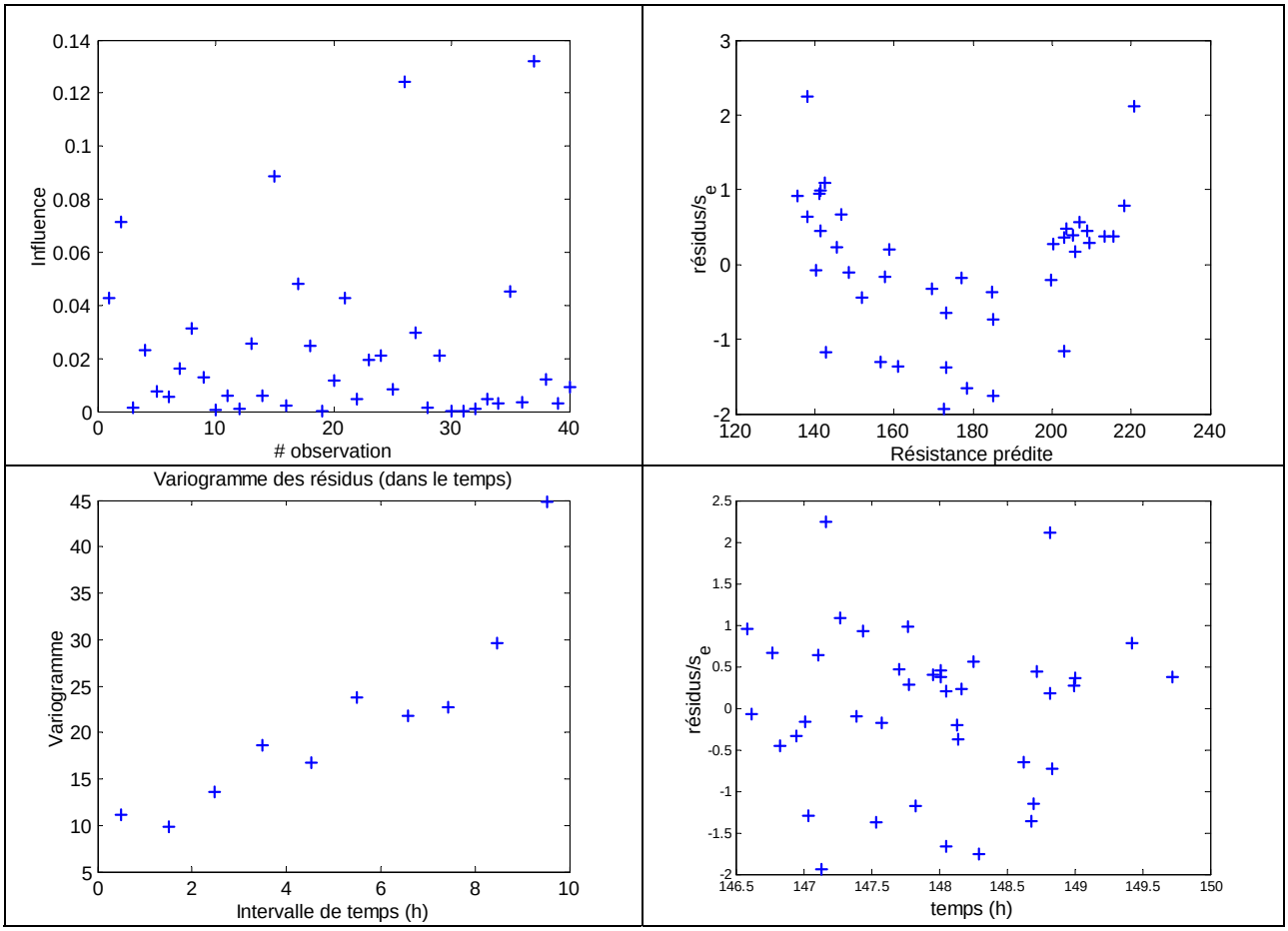
	Classement $\hat{Y}$ <0.5	Classement $\hat{Y}$ >0.5
Vrai Y=0	70	24
Vrai Y=1	23	84

2	*a) Les roches ayant subi l'altération liée à la minéralisation sont-elles enrichies ou lessivées en CaO et Na2O ?*
2	*b) Quel est le taux global de bonne classification de ce modèle ?*
4	*c) Supposons que vous prélevez une roche volcanique dans une région peu explorée. Quelle serait la probabilité que cette roche ait subi une altération du type rencontrée à proximité des gisements volcanogènes si sa composition est : Al2O3*=13%, CaO=1.5% et Na2O=2.5% ?

---

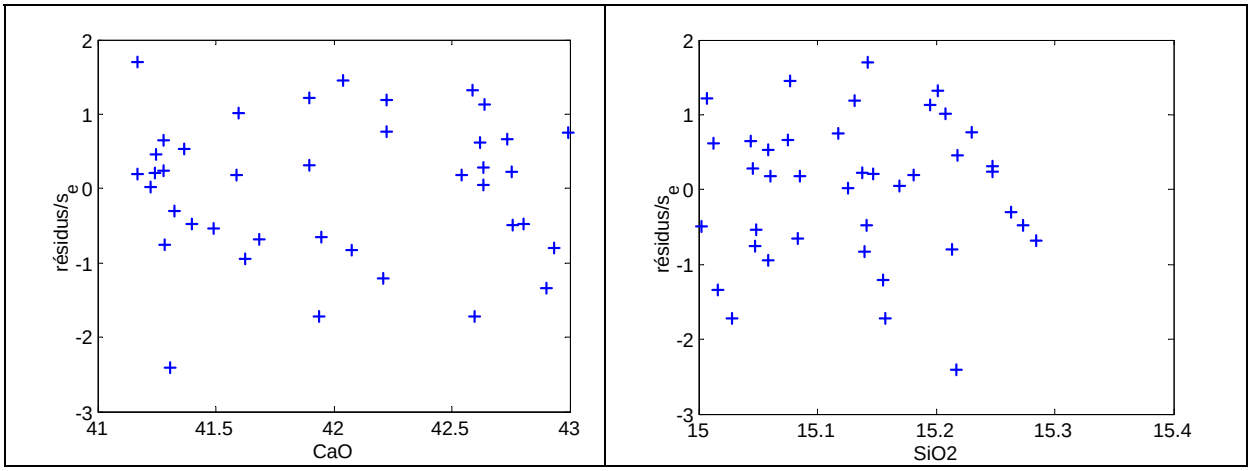
4. On veut prédire la résistance d'un ciment (kPa) à partir de la composition chimique du calcaire prélevé à la carrière. On établit un modèle de prédiction puis on calcule les résidus de ce modèle. Voici 8 graphes choisis des résidus :

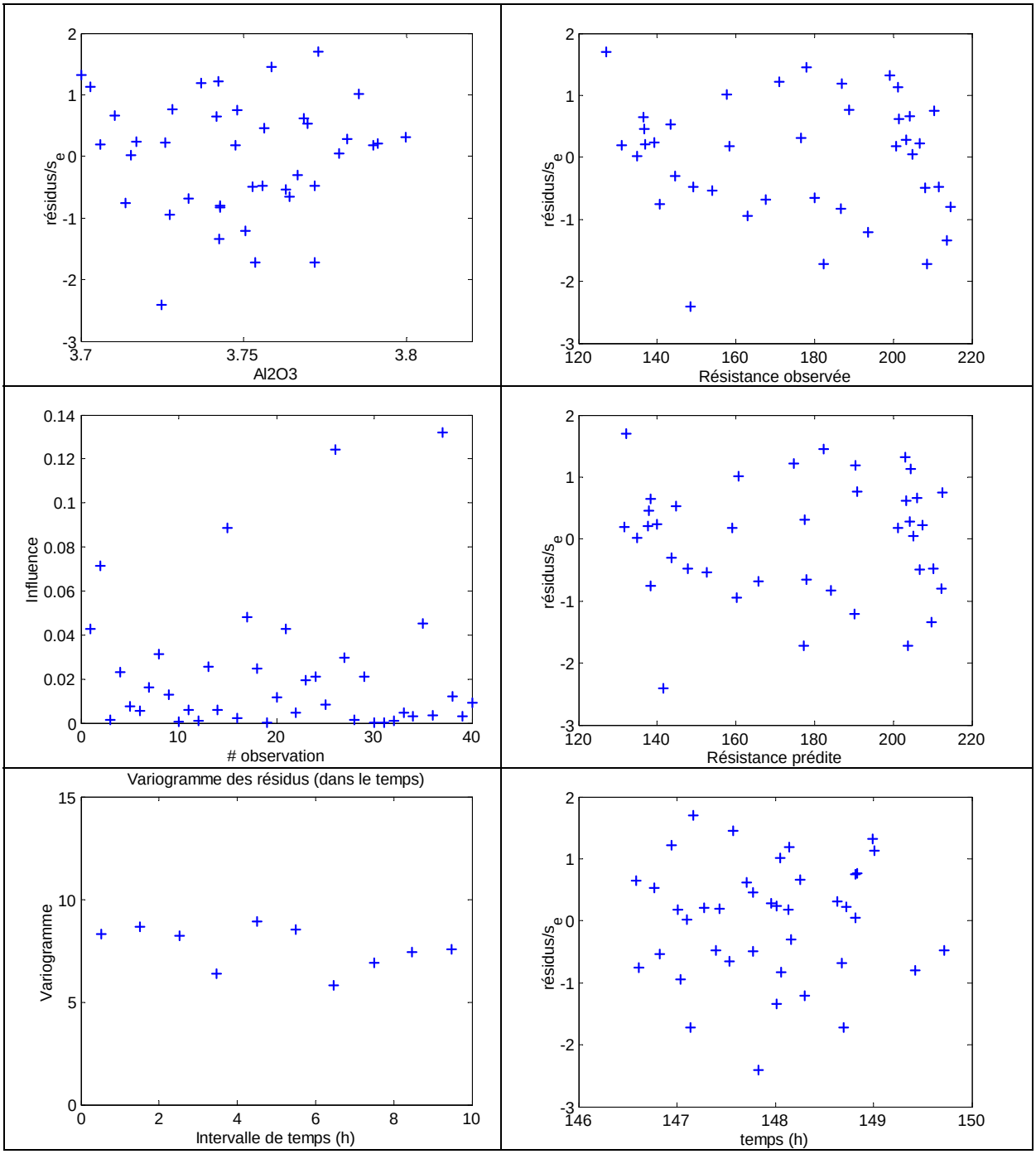




4 a) Identifiez tout comportement des résidus qui vous semble suspect en indiquant le problème suspecté.

On effectue une certaine correction pour améliorer le modèle et voici maintenant les graphes des résidus obtenus :





2	*b) La correction apportée a-t-elle réussi à corriger le(s) problèmes(s) identifié(s) précédemment ? Discutez.*

Corrigé

1- a) signe négatif car la vitesse décroît avec la distance et augmente avec la charge utilisée
b) Y : ln(v)
X : colonne de 1, colonne avec $d/w^{0.5}$.
c) Y : ln(v)
X : colonne de 1, colonne avec ln(d) et colonne avec ln(w)

d) On a R^2=0.8 =1-SCE/SCT_m=1-2/SCT_m. On trouve SCT_m=2/(1-0.8)=10.
SCR_m=SCT_m-SCE=10-2=8.

Avec le modèle complet, On a $SCR_{m,c}$=0.9*10=9. SCE_c=10-9=1
Le test est donc (2-1)/1 / 1/(30-3) = 27 > $F_{1,27;.05}$=4.21. L'ajout est très significatif.

e) il suffit de combiner les 50 observations et de considérer le second modèle comme modèle réduit et d'ajouter 3 variables permettant une droite de régression différente pour chaque cas : I, I*ln(d) I*ln(w). On teste l'ajout de ces 3 coefficients simultanément par rapport au modèle précédent. Le test comprend au numérateur la différence des SCE entre modèle réduit et complet avec 3 degrés de liberté. Au dénominateur, c'est SCE du modèle complet avec maintenant 30-6=24 degrés de liberté.

f) le problème est lié à la détermination de « w ». Les 2 modèles de régression sont construits pour prédire ln(v) et non ln(w). Il ne s'agit donc pas de l,estimation la meilleure que l'on puisse faire de ln(w) avec les données disponibles. Il faudrait faire une régression de ln(w) en fonction de ln(v) et ln(d) pour prédire de façon optimale ln(w).

2- a) La première variable incluse est celle avec le + grand coefficient de corrélation simple soit ici le ciment siliceux (r=0.78). Pour savoir laquelle des deux autres entre dans la régression, il faut calculer la corrélation partielle de « def » vs qtz grossier et qrz fin étant donné Ciment siliceux contrôlé.

$r_{def,qtz\ grossier|ciment} = 0.28-(0.78*0.59)/((1-0.78^2)(1-0.59^2))^{0.5}= -0.36$

$r_{def,qtz\ fin|ciment} = 0.22-(0.78*-0.26)/((1-0.78^2)(1-0.26^2))^{0.5}= 0.70$

la 2^e variable qui entre est donc le qtz fin.

b) Après la 1ère variable, le R^2 vaut 0.78^2=0.61
Il reste 1-0.61=0.39 à expliquer. La corrélation partielle de 0.7 indique que 0.7^2 =0.49 de ce 0.39 sera expliqué. Au total R^2 vaudra donc : 0.61+0.49*0.39=0.80
(voir notes p.37, eq. 2.30)

c) On a R^2=0.8. On a aussi $Y'\hat{Y} = 0.063 = \hat{Y}'\hat{Y} = SCR$. Comme , on trouve $\overline{Y} = 0.047$, on trouve SCM=17*.047^2=.038. SCR_m=0.063-0.038=0.035. SCT_m=SCR_m/R^2=.035/0.8=.044 et SCT=.044+.038=.0.082. SCE =SCT-SCR=.082-.063=.019.
le tableau suivant résume la situation:

SC	valeur	d.l.
SCT	.082	17
SCM	.038	1
SCT_m	.044	16
SCR	.063	3
SCR_m	.035	2

SCE	.019	14

d) On a R^2=0.61 avec la 1ère variable. donc $SCR_{m,r}=R^2SCT_m$=0.61*.044=.027.

Test d'ajout : ((.035-.027)/1)/(.019/14)=5.9 à comparer à une $F_{1,14}$=4.6 L'ajout est significatif.

3- a) plus il y a de CaO et de Na2O dans la roche plus faible est $\hat{W}$ et donc plus faible est la probabilité que l'altération se soit produite. On en conclut que l,altération est marquée par un lessivage (perte) en Na2O et CaO.

b) 154/201=76.6%

c) On calcule $\hat{W}$ =5.68-0.25*13-0.38*1.5-0.41*2.5=0.835
$\hat{Y}$ =exp($\hat{W}$)/(1+exp($\hat{W}$))=0.70

4- a) i. On note une structure en arche sur le diagramme résidus vs CaO (inclure $Cao^{0.5}$ ou CaO^2?
ii. on note une structure en arche sur le diagramme résidus vs valeurs prédites
iii. on note que le variogramme des résidus n'est pas un effet de pépite pur dans le temps. Il y a une composante temporelle manquante.

Note : le graphe résidus vs valeurs observées n'est pas comme attendu mais on ne doit pas l'utiliser pour établir le diagnostique.

b) tout est beau maintenant, le variogramme des résidus est un effet de pépite pur et les structures en arche sont disparues. Aucune observation n'a d'influence trop grande.

GLQ3402 -- Examen de mi-session

Jeudi le 12 juin 2003
8h30 à 11h00

Toute documentation permise.
Calculatrice permise

L'examen comporte 9 questions totalisant 30 points. Chaque réponse doit être accompagnée des calculs et justifications appropriés.

Les questions valent dans l'ordre : 3, 7, 3, 3, 3, 4, 2, 3, 2

Points

3 1- Dans Excel, la fonction « Growth » (ou « Croissance » en français) permet d'estimer le modèle :

$$\hat{Y} = b_0 b_1^x$$

a) Indiquez comment vous pourriez « linéariser» ce modèle et ainsi obtenir des estimés pour b_0 et b_1 avec un programme de régression linéaire. Indiquez clairement le vecteur « y » et la matrice « x » qui seront soumis au programme de régression, les coefficients obtenus par la régression et le lien avec les coefficients recherchés.

b) Les prédictions $\hat{Y}$ *obtenues avec ce modèle minimiseront-elles la somme des carrés des erreurs (si l'on définit « e » comme* $e = Y - \hat{Y}$ *) ? Justifiez.*

---

7 2- On effectue une régression pas à pas (avant) avec 100 observations. On veut expliquer une variable Y avec au plus 4 variables X (modèle avec constante). Les 2 premières variables à être incluses dans la régression sont X_2 et X_4. À cette étape, le R^2 vaut 0.8 et SCE vaut 1000. L'on vous fournit la matrice de **variances-covariances partielles** (étant donné l'effet de X_2 et X_4 fixé) suivante :

	Y	X_1	X_3
Y	10	14.23	28.3
X_1	14.23	25	33.54
X_3	28.3	33.54	125

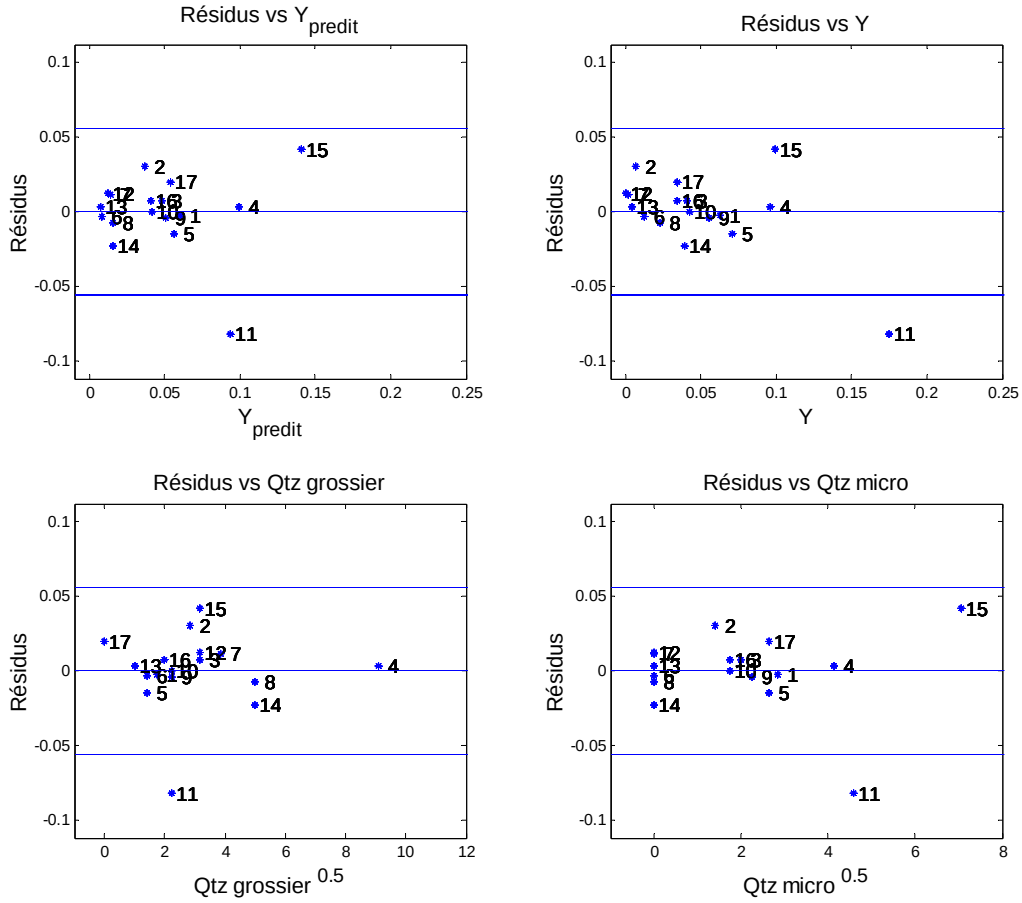
a) Quelle sera la prochaine variable à être sélectionnée par la procédure pas à pas avant dans le modèle de régression? (Aide : quel est le lien entre covariance partielle et corrélation partielle?)
b) Quelle est le coefficient de la régression associé à cette variable?
c) Les coefficients de régression associé à X_2 et X_4 seront-ils modifiés suite à l'inclusion de cette nouvelle variable?
d) Que devient le R^2 suite à l'inclusion de cette variable?
e) L'ajout dû à cette variable est-il significatif? Faites le test requis en indiquant clairement les degrés de liberté associés. (Aide, calculez d'abord SCT_m).

3 3- On a déterminé la porosité d'échantillons de résidus miniers provenant de 2 sites différents.

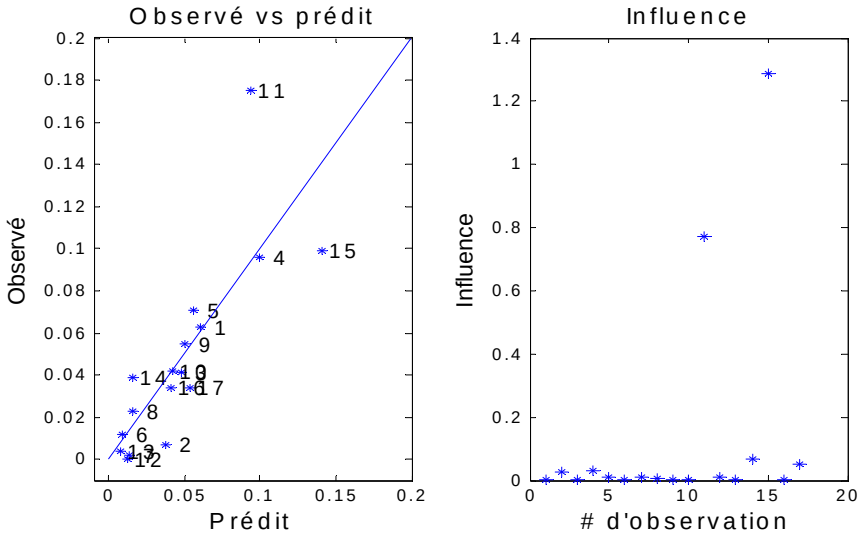
Expliquez comment on peut utiliser un programme de régression pour tester l'égalité des moyennes des porosités des 2 sites. (Aide : on doit utiliser le test d'ajout)

---

3 4- On a mesuré la déformation de 17 éprouvettes de béton après une cure humide d'un an. On a aussi déterminé le % de quartz micro-cristallin (chert et calcédoine) et le % de grains de quartz plus grossiers pour chaque éprouvette. Ces déterminations ont été réalisées par comptage microscopique. L'objectif de la régression est de prédire la déformation après un an en utilisant les % de quartz comme variables prédictives. Les graphiques suivants montrent les résidus obtenus en fonction de certaines variables. Les limites correspondant à + ou – 2 *CME$^{0.5}$ sont aussi illustrées :



Le graphique suivant montre le graphe des valeurs observées vs prédites et celui de l'influence des observations :



Les résultats présentés vous semblent-ils conformes à ce qui est attendu ? Justifiez votre réponse.

---

3 5- Toujours avec le jeu d'éprouvettes de béton de la question précédente, on a appliqué la norme ACNOR qui indique qu'un béton doit être rejeté s'il subit une déformation supérieure à 0.04. Ceci permet de définir deux groupes de béton, ceux ayant réussi et ceux ayant échoué le test. On applique une régression logistique à ces données et l'on obtient le modèle suivant :

$$\hat{W} = -3.6 + 0.07 * Qtz\ \ grossier(\%) + 0.8 * Qtz\ \ micro(\%)$$

a) Supposons qu'un béton contienne 5% de quartz grossier. Quelle quantité de quartz micro-cristallin peut-il posséder avant que l'on ne doive rejeter le béton (si l'on se base sur l'équation de prédiction logistique)?

b) Le béton contient 20% de quartz grossier et 8% de quartz micro-cristallin. Quelle est la probabilité qu'il échoue le test de la cure humide?

---

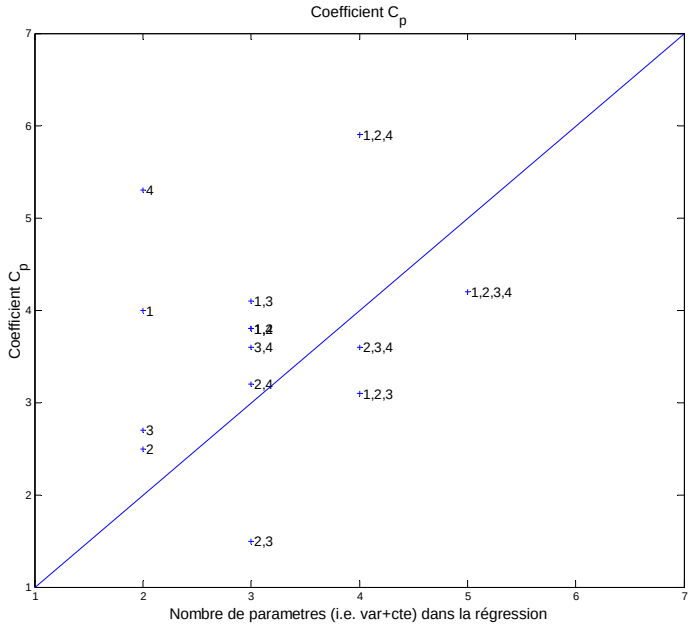
4 6- Soit la matrice de corrélation simple suivante obtenue avec les données géochimiques de l'Abitibi utilisées lors du TP-4.

	SiO_2	Al_2O_3	FeO	MgO	CaO	Na_2O	K_2O	TiO_2
SiO_2	1.00	-0.84	-0.65	-0.54	-0.80	-0.48	0.31	-0.84
Al_2O_3	-0.84	1.00	0.31	0.27	0.61	0.51	-0.08	0.65
FeO	-0.65	0.31	1.00	0.59	0.31	-0.02	-0.37	0.63
MgO	-0.54	0.27	0.59	1.00	0.32	-0.21	-0.27	0.39
CaO	-0.80	0.61	0.31	0.32	1.00	0.44	-0.25	0.64
Na_2O	-0.48	0.51	-0.02	-0.21	0.44	1.00	-0.44	0.43
K_2O	0.31	-0.08	-0.37	-0.27	-0.25	-0.44	1.00	-0.35
TiO_2	-0.84	0.65	0.63	0.39	0.64	0.43	-0.35	1.00

a) Calculez la corrélation partielle entre FeO et MgO quand l'on fixe l'effet du SiO_2.
b) On effectue la régression de « FeO expliqué par SiO_2 » et la régression de « MgO expliqué par SiO_2 ». Quelle est la corrélation simple entre les résidus de ces deux régressions?
c) Quelle est la corrélation simple entre le résidu de la régression « FeO expliqué par SiO_2 » et le SiO_2?

---

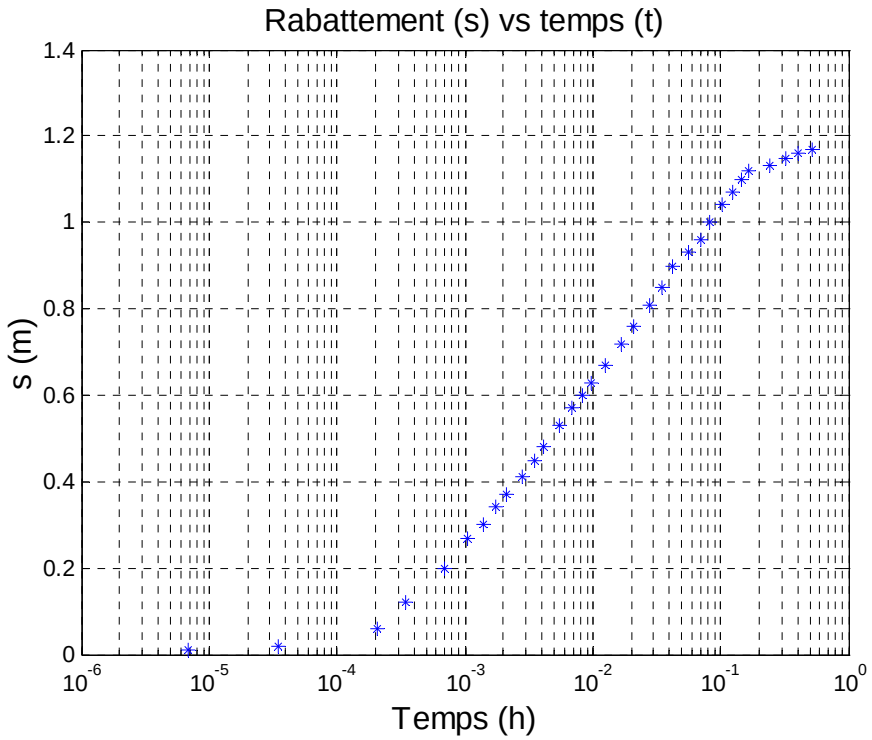
2 7- Soit le diagramme suivant illustrant le coefficient C_p obtenu pour différents sous-ensembles de variables.



Quel ensemble de variables ce graphique vous suggère-il retenir? Justifiez.

---

3 8- Dans un aquifère à nappe confinée, homogène, de grande extension latérale et d'épaisseur constante, l'on dispose d'un piézomètre ayant une crépine sur toute l'épaisseur de l'aquifère. On mesure le rabattement (s) en fonction du temps en vu d'estimer la transmissivité (T) et le coefficient d'emmagasinement (S) à l'aide de la méthode de Cooper-Jacob qui consiste à ajuster une droite sur la partie linéaire du graphe s vs ln(t). Le graphique suivant montre le rabattement en fonction du temps.



Suggérez un outil vu au cours qui pourrait permettre d'identifier et d'éliminer de façon itérative les points qui s'éloignent trop de la partie droite de la courbe.

---

2 9- On veut utiliser une sonde mesurant la conductivité électrique en continu sur un convoyeur pour identifier les fragments minéralisés et les fragments non-minéralisés en vu de les séparer. La teneur économique pour le gisement est 0.5% Ni équivalent (le seuil définissant « minéralisé »). On définit le Ni équivalent comme Ni+0.8*Cu. Vous avez mesuré la teneur chimique en Ni et en Cu sur 100 fragments et avez mesuré sur ces mêmes fragments la conductivité électrique. Vous notez une relation linéaire assez forte entre le log(conductivité) et le Ni équivalent.

Quel modèle de régression utiliserez-vous pour rencontrer les objectifs de l'étude?

Bonne chance,

Denis Marcotte

Corrigé :

1- a) En prenant les logs, on obtient : $ln(\hat{Y}) = ln(b_0) + ln(b_1) * x$. On a donc une régression linéaire avec constante de ln(Y) prédit par x. On obtient c_0 et c_1 de cette régression et $b_0=\exp(c_0)$ $b_1=\exp(c_1)$.

b) Comme la régression s'effectue dans un espace transformé, elle minimise la somme du carré des erreurs de la variable transformée ln(Y) et non du Y lui-même. Si l'on veut minimiser la somme du carré des erreurs, il faut effectuer une régression non-linéaire.

2- a) On cherche la variable avec la plus forte corrélation partielle avec Y. On calcule pour
X_1 : $14.23/(10*25)^{0.5}=0.9$
X_3 : $28.3/(10*125)^{0.5}=0.8$
On doit donc inclure X_1.

b) Le coefficient de la régression est : 14.23/25=0.569
c) Oui bien sûr puisque b est obtenu par $(X'X)^{-1}X'Y$ et que la matrice X change de dimension.
d) Le nouveau R^2 peut être calculé par la relation 2.30 donnée en page 37.
$R^2=0.8+0.9^2*(1-0.8)=0.962$
e) Le SCE avec X_2 et X_4 vaut 1000, on déduit que $1-R^2=0.2=1000/SCT_m$ et $SCT_m=5000$.
Après avoir inclus X_1, on a $1-R^2=0.038=SCE(c)/5000$ donc SCE(c)=190.
Le test d'ajout est donc $F_{calculé}=\dfrac{(1000-190)/1}{190/(100-4)}=409.3$, ce qui est très supérieur à la valeur

$F_{table,.1,96,05}=3.94$

3- On forme un premier modèle réduit $Y=b_o+e$. On forme un second modèle complet : $Y=b_0+b_1*I+e$ où I est une variable indicatrice prenant la valeur 0 pour un site et 1 pour l'autre. Ce modèle permet d'avoir deux moyennes différentes comme estimation pour les deux sites alors que le modèle réduit ne permets qu'une seule moyenne. Il ne reste qu'à tester le caractère significatif de l'ajout. Si c'est significatif alors les moyennes diffèrent, significativement, sinon, les moyennes peuvent être considérées égales.

4- On note que l'observation 15 montre une trop grande influence. De plus l'observation 11 montre aussi une grande influence et son résidu est à l'extérieur de l'intervalle de confiance calculé avec $CME^{0.5}$. Cette observation est donc aussi très suspecte. Il faudrait possiblement refaire l'analyse sans ces 2 observations.

5- a) Le seuil de classification correspond à $\hat{W}=0$. On calcule donc
% qtz micro<(3.6-0.07*5)/0.8=4.06%.
b) On calcule $\hat{W}=-3.6+0.07*20+0.8*8=4.2$
$\hat{Y}=exp(4.2)/(1+exp(4.2))=0.985$

6- a) On calcule $r_{Feo,MgO|SiO2} = \frac{0.59 - (-0.65) * (-0.54)}{\left((1 - 0.65^2)(1 - 0.54^2\right)^{0.5}} = 0.37$

b) Par définition, elle est égale à la corrélation partielle calculée en a)

c) Les résidus d'une régression sont non-corrélées aux variables « x » entrant dans la régression. Donc cette corrélation est 0.

7- Tous les ensembles ayant les variables x_2 x_3 sont sous la diagonale. Parmi ceux-ci le couple (2,3) est celui le plus sous la diagonale. Il s,agit probablement du meilleur sous-ensemble à retenir.

8- La notion d'influence d'une observation peut permettre de détecter les observations s'écartant de la droite. En enlevant les observations les plus influentes et en faisant le suivi du R^2 obtenu, on pourra déterminer un sous-ensemble de points situés sur une droite.

9- Le modèle aura la forme : Ni équivalent=b_0+b_1 log(conductivité électrique)+e

GLQ3402 -- Examen de mi-session

Lundi le 3 juin 2002
13h00 à 15h30

Toute documentation permise.
Calculatrice permise

L'examen comporte 6 questions totalisant 100 points. Chaque réponse doit être accompagnée des calculs et justifications appropriés.

Les questions valent : 1-20, 2-30, 3-20, 4-10, 5-10, 6-10

Points

20 1- On désire modéliser l'écoulement régional d'un aquifère de surface situé sur la rive sud de Montréal. Pour ce faire, on effectue le relevé de 60 piézomètres en un court laps de temps au mois d'avril (moment où la nappe est approximativement à son niveau maximal). On note les coordonnées *x, y* donnant la localisation en plan des piézomètres, *z* l'élévation du sol au piézomètre et *h* la charge hydraulique. On tente d'ajuster par régression divers modèles et l'on obtient les résultats suivants :

Modèle	Équation	SCE (m^2)
A	$h(x,y)=b_0+e$	70
B	$h(x,y)=b_0+b_1x+b_2y+e$	20
C	$h(x,y)=b_0+b_1x+b_2y+b_3z+ e$	10
D	$h(x,y)=b_0+b_1z+ e$	13

a) Quel modèle devrait-on retenir? Indiquez vos calculs et faites les tests nécessaires.

b) Que vaut le R^2 de ce modèle ?

On refait le même relevé, mais cette fois au mois de septembre. Pour la suite, on suppose que le modèle C est le modèle retenu.

c) Bien que le niveau moyen de la nappe en septembre soit certes inférieur à celui du mois d'avril, comment pourriez-vous vérifier néanmoins que la direction régionale d'écoulement demeure la même (d'un point de vue statistique) pour ces deux périodes de l'année ? Indiquez les modèles utilisés et les degrés de liberté associés au test. Aide : partant du modèle C construisez un modèle imposant les mêmes coefficients de régression pour x et y aux deux jeux de données (avril et septembre). Comparez ce modèle à un modèle permettant des coefficients différents pour x et y selon le relevé.

30 2- Lorsque l'on estime les réserves d'une mine, il est important de tenir compte des variations de densité de la roche. Souvent ces variations de densité sont liées assez directement à la teneur du minerai, surtout lorsque le minerai est riche et constitué de sulfures plus denses que la roche mère. Conscient de ce fait, un ingénieur géologue d'une mine de Cu-Zn (Cu dans la chalcopyrite, Zn dans la sphalérite) a mesuré avec précision la densité de 100 échantillons représentatifs de la mine. Il a de plus effectué l'analyse géochimique de ces mêmes échantillons. Le tableau suivant présente une partie des résultats qu'il a obtenus (d_i : densité de l'observation i (sans unité); $CuZn_i$: Cu+Zn pour l'observation i, en %) :

Statistique	Valeur
Moy(d)	3.1
Var(d)	17
Moy(CuZn)	4.2 %
Var(CuZn)	280 $\%^2$
Cov(d,$CuZn_i$)	28 %
SCT_m	1700

a) Utilisant ces résultats, et supposant que la densité du minerai est liée linéairement aux teneurs de Cu+Zn, donnez l'équation de prédiction de la densité du minerai.

b) Quelles sont les unités associées aux coefficients b_0, et b_1 de la régression ?

c) Que vaut le coefficient de corrélation simple ?

d) Que vaut le coefficient de détermination R^2 ?

e) Que vaut CME ?

f) Un minerai montre une teneur de 4% Cu et 5% Zn. Quelle est la valeur prédite de la densité?

g) En additionnant les teneurs en Cu et Zn pour effectuer sa régression, quelle hypothèse a été faite implicitement par l'ingénieur?

---

20 3- Dans une cimenterie l'objectif principal est de fournir un produit homogène rencontrant le plus possible les spécifications propres à chaque type de ciment produit. Une étape cruciale de la fabrication du ciment est le mélange des matériaux bruts (calcaire, argile, fer et silice) en des proportions qui sont constamment et automatiquement ajustées par un logiciel de contrôle. Comme input à ce logiciel de contrôle, il est important de fournir régulièrement et précisément les teneurs en 8 éléments majeurs des différents matériaux bruts. Une cimenterie utilise un appareil Gammametrics pour mesurer ces teneurs en continu sur les différents convoyeurs. Cet appareil utilise différentes sondes géophysiques et un modèle théorique (boîte noire) pour convertir les mesures en teneurs équivalentes. Afin de calibrer les teneurs fournies par le Gammametrics, on analyse 50 échantillons sur chacun des quatre convoyeurs et l'on fait la régression élément par élément en utilisant les valeurs correspondantes du Gammametrics. On a ainsi 32 régressions différentes (8 variables x 4 convoyeurs). Pour chaque régression, les variables explicatives sont l'ensemble des 8 teneurs Gammametrics.

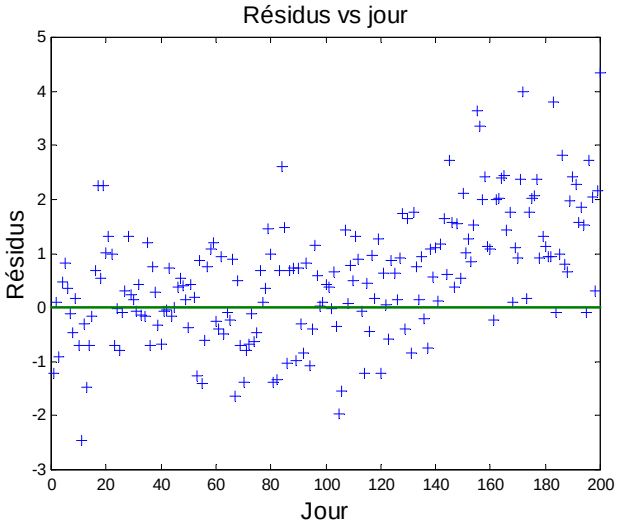
Ainsi, pour le SiO_2 pris sur le convoyeur de calcaire, l'on obtient par une procédure pas à pas le modèle suivant (les unités des variables sont des %):

$SiO_{2,chimique}$= -0.5+0.94* $SiO_{2,gamma}$+0.08*CaO_{gamma} +e

Le R^2 de ce modèle est 0.95. Lorsqu'on effectue la régression (sans constante) avec uniquement le $SiO_{2,gamma}$ on obtient un R^2 de 0.91 et le coefficient b_1 de la régression vaut 1. On a de plus SCT= 7000%^{2} et la moyenne $\overline{SiO}_{2,chimique}$ =10%

a) Selon ces résultats, devrait-on fournir au logiciel de contrôle la teneur en SiO_2 obtenue du Gammametrics ou la teneur en SiO_2 estimée par la régression ? Justifiez.

Supposons qu'en a) on ait finalement retenu le modèle de régression avec SiO_2 et CaO. L'ingénieur de procédé prend la précaution d'échantillonner 1 fois par jour chaque convoyeur et d'effectuer l'analyse chimique. Il compare les teneurs observées avec celles prédites avec le Gammamétrics et le modèle de régression (qui demeure inchangé dans le temps). Voici ce qu'il observe au bout d'un certain temps pour le SiO_2 (sur le convoyeur de calcaire) :



b) Que conseillez-vous de faire ? Justifiez.

---

10 4- Avec les données de « solgele » on a observé que la procédure pas à pas sélectionnait dans l'ordre les variables 11 (masse vol. totale), 1 (% poids de diesel dans les fluides) et la variable 12 (masse vol. sèche). Pourtant lorsque l'on examine le diagramme binaire de la résistance vs le % diesel, on ne note pas de relation très forte alors que la relation est assez forte entre la résistance et la masse volumique totale. Ces observations se retrouvent aussi dans la matrice de corrélation simple :

	Résistance	Diesel	Masse tot.	Masse sèche
Résistance	1.00	-0.20	0.97	0.95
Diesel	-0.20	1.00	-0.07	-0.16
Masse tot.	0.97	-0.07	1.00	0.98
Masse sèche	0.95	-0.16	0.98	1.00

Toutefois, la matrice de corrélation partielle obtenue en fixant la masse volumique totale est :

	Résistance	Diesel	Masse tot.	Masse sèche
Résistance	1.00	-0.58	0.00	-0.17
Diesel	-0.58	1.00	0.00	-0.53
Masse tot.	0.00	0.00	0.00	0.00
Masse sèche	-0.17	-0.53	0.00	1.00

a) À la 2^{e} étape, pourquoi le diesel a-t-il été sélectionné par l'algorithme pas à pas de préférence à la masse sèche?

b) Avancez une explication géologique plausible pour le fait que la corrélation simple entre la contrainte et la masse sèche soit très fortement positive et qu'elle devienne négative si l'on fixe la masse volumique totale.

---

10 5- *Parmi les 4 modèles suivants, un seul peut, après transformation, être traité par régression linéaire. Indiquez quel est ce modèle et donnez son expression après transformation.*

i. $Y = b_0 + b_1X_1 + b_2X_2^{b_3} + e$

ii. $Y = b_0X_1^{b_1}e$

iii. $Y = b_0X_1^{b_1} + e$

iv. $Y= b_0 + b_1Cos(b_2X_1)+e$

---

10 6- *Dans une régression linéaire multiple, à quelle somme de carré le produit Y'Xb est-il égal ? Démontrez.*

Corrigé :

1- a) Test d'ajout ; on peut tester D par rapport à A puis C par rapport à D.
D vs A : F calculé : (57/1)/(13/58)=254 Fortement significatif
C vs D : F calculé : (3/2)/(10/56)=8.4 > 3.16=F(2,56 ;0.95) Significatif on conserve C

b) SCTm=70, SCE=10, SCRm=70-10=60, R^2=60/70=0.86.

c) Il suffit de mettre les 2 jeux de données ensemble et de tester le modèle

F : $h(x,y)=b_0+b_1x+b_2y+b_3z+b_4I+b_5Iz+e$
vs
G: $h(x,y)=b_0+b_1x+b_2y+b_3z+b_4I+ b_5Iz +b_6Ix+b_7 Iy+e$

En testant l'ajout de G vs F, on teste l'égalité à 0 de b_5 et b_6. Si l'on accepte l'hypothèse alors les directions régionales d'écoulement sont les mêmes, sinon il y a changement significatif. Le test comporte 2 degrés de liberté au num,érateur et 60-8=52 au dénominateur.

2- a) $b_1=s_{xy}/s_x^2$= 28/280=0.1 $(\%)^{-1}$
b_0=3.1-0.1*4.2=2.68

b) b0 est sans unité et b1 est en $\%^{-1}$

c) $r=28/(17*280)^{0.5}$=0.4058

d) R^2= 0.4058^2=0.1647

e) SCRm=R^2*SCTm=1700*0.1647=280
SCE=1700-280=1420
CME=1420/(100-2)=14.5

f) valeur prédite: 2.68+9*0.1=3.58%

g) que les coefficients de la régression pour ces 2 variables étaient égaux et donc que la densité de ces 2 minéraux (sphalérite et chalcopyrite étaient égales)

3-
a) Test d'ajout modèle réduit : sans constante, modèle complet avec constante et CaO.

SCTm=7000-50*10^2=2000.
SCRm(r)=0.91*2000=1820
SCRm(c)=0.95*2000=1900
SCE(c)=2000-1900=100

F calculé : $\frac{(1900-1820)/2}{100/(50-3)}=18.8$ ce qui est très supérieur à la valeur d'une F(2,47 ;0.05)=3.19.
On doit donc utiliser le modèle corrigé par la régression.

b) Les résidus montrent une tendance croissante à partir du 120^e jour environ. On devrait recalibrer l'appareil en effectuant une nouvelle régression sur un bon nombre d'observations. On pourrait aussi

envisager mettre à jour le modèle en utilisant les « n » dernières données disponibles où « n » pourrait-être déterminé par une étude de type validation croisée.

4- a) la corrélation partielle est maximale avec le diesel. C'est donc cette variable qui doit entrer dans la régression.

b) en fixant la masse volumique totale, si l'on augmente la masse volumique sèche, alors la masse de fluides diminue dans l'échantillon et celui-ci devient moins résistant. Ayant moins de liquides, les grains de sol sont moins liés par la glace et la rupture arrive plus rapidement.

5- C'est le modèle ii. On prend le log : $\ln(Y)=\ln(b_0)+b_2 \ln(X_2) +\ln(e)$

6- $Y'Xb=Y'X(X'X)^{-1}X'Y=Y'MY=SCR$

GLQ3402 -- Examen de mi-session

Lundi le 4 juin 2001
13h00 à 15h30

Toute documentation permise.
Calculatrice permise. Téléphones cellulaires interdits.

L'examen comporte 6 questions totalisant 100 points. Chaque réponse doit être accompagnée des calculs et justifications appropriés.

Les questions valent : 1-10, 2-21, 3-16, 4-20, 5-13, 6-20

Points

10 1- Dans un modèle de régression linéaire $Y=b_0+b_1X+e$, on sait que $b_0=\overline{Y}-b_1\overline{X}$.

Utilisez ce fait pour modifier le modèle initial et montrer ainsi que b_1 *peut être obtenu directement d'une régression sans constante.*

---

21 2- Pour un aquifère à nappe captive sans recharge et soumis à un pompage, on atteint un état d'équilibre liant le rabattement (s) (i.e. diminution de la charge par rapport à l'état initial) à la distance au puits de pompage (r). On a alors: $s=b_0+b_1\ln(r)+e$. Le coefficient b_1 est lié à la transmissivité par: $b_1=\frac{-Q}{2\pi T}$ où Q est le débit pompé ($100 cm^3/s$) et T est la transmissivité. On a disposé 3 piézomètres selon une direction en s'éloignant du puits et l'on a observé:

Piézomètre #	Rabattement s (cm)	Distance au puits r (cm)
1	203	300
2	108	500
3	42	800

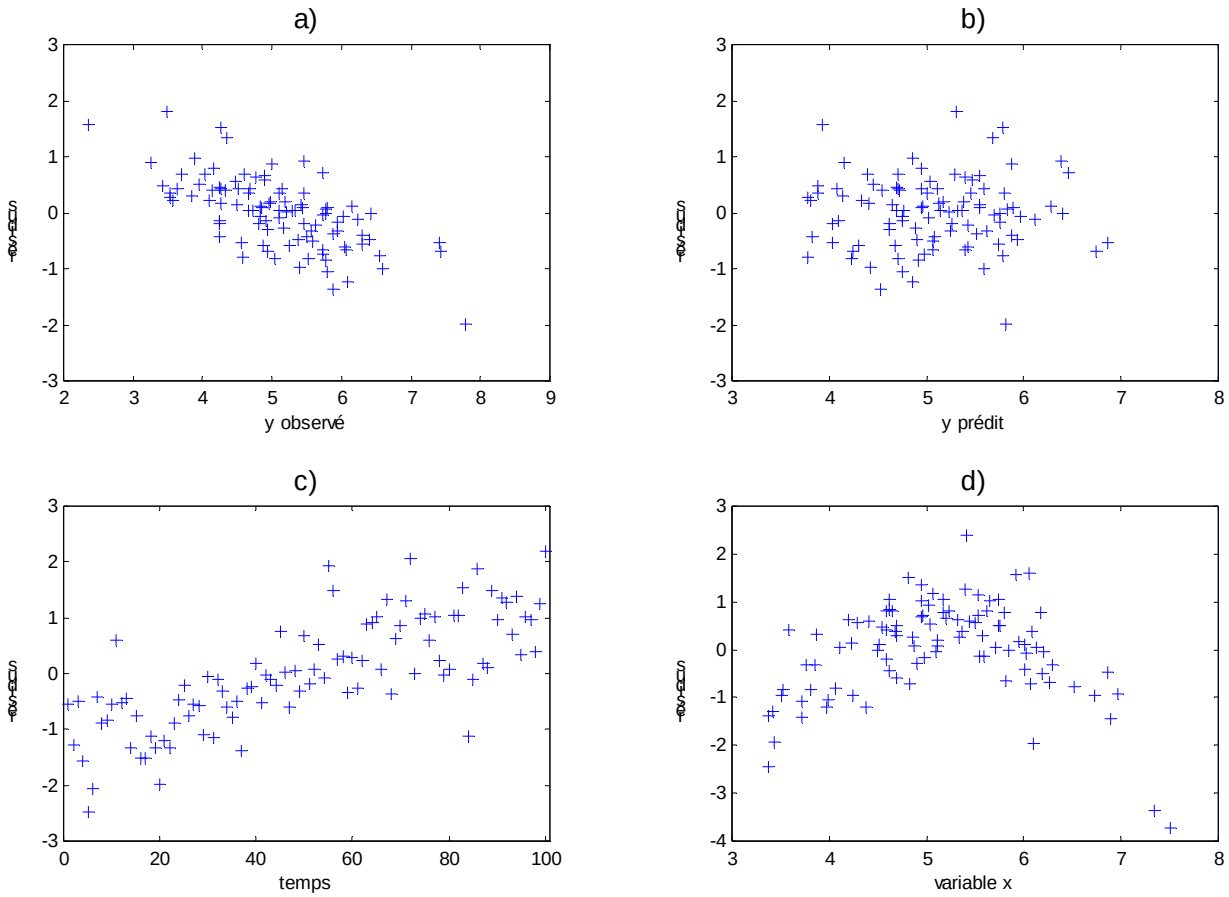
On calcule aussi les quantités suivantes:

$\sum s_i=362$	$\sum r_i=1600$	$\sum \ln(r_i)=18.6$	$\sum (s_i-\bar{s})(r_i-\bar{r})=-39567$
$\sum s_i^2=56342$	$\sum r_i^2=980000$	$\sum (\ln(r_i))^2=115.8$	$\sum (s_i-\bar{s})(\ln(r_i)-\overline{\ln(r)})=-78$
$\sum (s_i-\bar{s})^2=12661$	$\sum (r_i-\bar{r})^2=126667$	$\sum (\ln(r_i)-\overline{\ln(r)})^2=0.48$	SCE=3.19

a) Calculez le coefficient b_1 *de la régression, et déduisez la transmissivité de l'aquifère exprimée en* cm^2/s
b) Quel est l'intervalle de confiance de niveau 95% pour la transmissivité de l'aquifère?
c) On décide d'implanter 3 autres piézomètres aux mêmes distances mais dans une direction orthogonale à la direction définie par les 3 premiers. Indiquez comment on devrait utiliser la régression pour pouvoir tester si la transmissivité est la même selon les deux directions.

---

16 3- Le graphe suivant montrent les résidus d'une régression en fonction de diverses variables. Indiquez dans chaque cas si le résultat est suspect ou non. Si vous trouvez le résultat suspect, indiquez ce que vous pourriez faire pour tenter d'améliorer le modèle.



20 4- Utilisant des données similaires à celles du TP-2, on utilise une procédure pas à pas pour inclure les variables les plus significatives. À la 1ère itération on retient la bentonite et l'on obtient un R^2 de 0.60. On calcule ensuite les corrélations partielles en fixant l'effet de la bentonite et l'on obtient les corrélations partielles suivantes pour le $\log_{10}(k)$:

Corrélation partielle de $\log_{10}(k)$ avec porosité	0.04
…avec saturation	0.47
…avec fines	-0.42
…avec (porosité*saturation)/100	0.56

Les variances partielles des différentes variables sont:

Variance partielle de porosité	79.5
… de saturation	259.8
… de fines	31.6
…de (porosité*saturation)/100	39.4
…de $\log_{10}(k)$	0.49

a) *Quelle serait la 2e variable à inclure dans la régression?*
b) *Quel est le coefficient de la régression associé à cette variable? (Aide: dans le cas où l'on a 1 seule variable X, le coefficient de régression s'obtient de la covariance et des variances de X et Y; la covariance entre X et Y peut s'obtenir de la corrélation et des variances).*
c) *Donnez la valeur du nouveau R^2 après l'ajout de la variable.*
d) *Le fait d'inclure cette nouvelle variable modifie-t-il le coefficient de régression que l'on avait auparavant pour la bentonite? Justifiez.*

---

13 5- On désire estimer la moyenne d'un relevé gravimétrique par un modèle de type "trend-surface" afin de filtrer la dérive pour faciliter l'analyse par approche fréquentielle. On envisage 3 modèles possibles pour cette moyenne:

(1) Gravi=b0+e
(2) Gravi=b0+b1*x+b2*y+e
(3) Gravi=b0+b1*x+b2*y+b3*x^2+b4*x*y+b5*y^2+e

On a 30 observations.

Les sommes des carrés des erreurs des différents modèles valent:

Modèle	SCE ($mgal^2$)
(1)	9000
(2)	3000
(3)	2800

a) Quel modèle devrait-on retenir pour estimer la moyenne spatiale de l'anomalie gravimétrique? Faites les tests requis au niveau 5%.
b) Donnez les unités des coefficients de la régression si x et y sont en m et Gravi en mgal.
c) Expliquez pourquoi cette équation ne devrait pas être utilisée pour produire des cartes gravimétriques.

---

20 6- Les bétons sont souvent sujets à détérioration en raison de la réaction chimique entre les granulats du béton et les alcalis du ciment qui forme un gel de silice entraînant le gonflement du béton. C'est ce gonflement qui crée des contraintes et amène l'éclatement du béton et l'apparition de fissures. Il existe une norme ACNOR pour les bétons spécifiant que ceux-ci ne doivent pas gonfler de plus de 0.04% après 1 an en cure humide. On envisage prédire le respect ou non-respect de la norme en se basant sur une description pétrographique minutieuse des phases de silice présentes dans l'agrégat. On dispose de 17 agrégats soumis à la norme ACNOR et dont le % des phases minéralogiques de quartz (et/ou silice amorphe) ont été mesurées et réparties en fractions grossières et fines.

On identifie le modèle logistique suivant (Y=0 test ACNOR réussi, Y=1 test ACNOR échoué):

$$\hat{W}=-3.91+0.09*(\%\text{quartz grossier})+0.81*(\%\text{quartz ou silice à grain fin})$$

Le log(maximum de vraisemblance) de ce modèle est –5.39. Le log(maximum de vraisemblance) du modèle ayant uniquement la constante est –11.75.

Le tableau de classement est

	Classement selon le modèle	
	Devrait réussir	Devrait échouer
Norme ACNOR réussie	8	1
Norme ACNOR échouée	1	7

a) La régression est-elle significative (niveau 5%)?
b) Quelle est la forme de silice la plus susceptible de causer le gonflement?
c) Quelle quantité de quartz à gain fin peut-on inclure dans un agrégat avant de devoir déclarer celui-ci impropre à son utilisation dans le béton (supposez qu'il n'y a pas de quartz grossier)?
d) Discutez de l'utilisation possible de cette régression d'un point de vue pratique (comment l'utiliseriez-vous? Avez-vous des suggestions à faire?) .
e) Un béton ayant 10% de quartz grossier et 2% de quartz à grain fin possède quelle probabilité de réussir le test ACNOR selon ce modèle?

Bonne chance

Denis Marcotte

Corrigé:

1- On peut écrire $Y-\overline{Y}=b_1(X-\overline{X})+e$ qui est un modèle sans constante.

2- a) b1=-78/0.48=-162.5 cm

$T=\frac{-Q}{2\pi b_1}$= -100/(2*pi*-162.5)=0.098 cm^2/s

b) Var(b1)=Var(e)/0.48 (voir notes p.13)

Var(e)=SCE/1=3.19
Var(b1)=6.64

Intervalle su r b1 : +- t(1;0.95) se= +- 12.7*2.58= =- 32.8

b1 compris entre [-195.3,-129.7] T entre [.081 , .122]

c) On ferait la même régression qu'en b) avec les 6 observations et on noterait le SCE correspondant. On ferait ensuite la régression en codant une des 2 directions et en utilisant le modèle:
s=b0+b1*ln(r)+b2*I*ln(r) avec I=1 si une des direction et 0 pour l'autre. On teste ensuite le caractère significatif de b1 ou, ce qui est la même chose, on fait le test d'ajout. Si on rejette le caractère significatif alors on peut conclure que les 2 directions montrent la même transmissivité, sinon il existe une différence significative.

3- a) La tendance linéaire provient d'un mauvais choix en abscisse. Il aurait fallu mettre la valeur prédite (comme en b). Il est normal d'observer un lien linéaire entre résidus et valeurs observées.

b) Normal
c) Il y a une tendance linéare dans le temps. Il faudrait inclure cette variable dans le modèle.
d) Il y a une tendance quadratique en fonction de la variable x. Il faudrait inclure, en plus de x une variable x^2 ou peut-être $x^{0.5}$ ou du moins une transformation qui permette de ramener le graphe à une bande homogène autour du niveau 0.

4- a) À chaque étape la variable entrant dans la régression est celle montrant la plus forte corrélation partielle. Ici ce serait donc la variable t=(porosité*saturation)/100

b) le coefficient de la régression sera: cov partielle(log(k),t) /(s2(t)) où s2(t) représente la variance partielle. Ici on a la corrélation partielle et l'on a : cov (log(k), t)=r(log(k),t)*s(log(k))*s(t) où s désigne l'écart-type (partiel).
Combinant ces 2 équations, on trouve b1= r (log(k),t) s(log(k))/(s(t)= $0.56*0.49^{0.5}/39.3^{0.5}$=0.063

c) R2 précédent 0.60.
Nouveau R2=ancien R2+(corrélation partielle)^2*(1-ancien R2) (voir équation 2.30 p. 35)
$R2=0.60+0.6^2*0.4=0.744$

d) Oui bien sûr. Les coefficients changent dès que l'on ajoute une nouvelle variable car ils sont données par b=inv(X'X)*X'Y. En ajoutant une variable on modifie X et donc b. Ainsi pour calculer le

coefficient mis à jour de la bentonite, il faudrait fournir les covariances partielles et variances partielles entre bentonite et log(k) quand on fixe l'effet de la variable t.

5- (2) vs (1) : $\frac{(9000-3000)/2}{3000/(30-3)}=27$ vs F(2,27;0.05)=3.35 Fortement significatif

(3) vs (2) $\frac{(3000-2800)/3}{2800/(30-6)}=0.57$ vs F(3,24,0.05)=3.01 Non-significatif. On retient la dérive linéaire.

b) Coefficient pour b0: mgal; b1: mgal/m; pour b2: mgal/m
c) Parce que la carte obtenue par "trend surface" peut donner des résultats insensés dès que l'on s'éloigne des données. De plus, elle ne fournit pas une interpolation exacte aux points des données et elle lisse beaucoup les détails des variations.

6 a) 2*(11.75-5.39)=12.72 vs Khi2 (2,0.05)=5.99 . Oui significatif.
b) nettement la silice à grain fin car son coefficient est 9 fois plus élevé.
c) on cherche le point où w=0. i.e. 3.91/0.81=4.8% de quartz à grain fin.
d) On pourrait s'en servir pour rejeter rapidement les agrégats problématiques. Pour les agrégats dont on prédit qu'ils ne gonfleront pas, il serait plus prudent d'effectuer la cure humide avant de l'accepter. Finalement, il serait intéressant de compléter cette étude par l'inclusion d'un plus grand nombre d'agrégats dans la régression.
e) w=-3.91+10*.09+2*0.81=-1.39
Prob d'échouer: exp(-1.39)/(1+exp(-1.39))=0.20.
Prob. Réussir=80%

GLQ3402 -- Examen de mi-session

Mercredi le 31 mai 2000
13h00 à 15h30

Toute documentation permise.
Calculatrice permise.

L'examen comporte 6 questions totalisant 100 points. <u>Chaque réponse doit être accompagnée des calculs et justifications appropriés.</u>

Points des questions : dans l'ordre : 10, 25, 5, 20, 20, 20

Points

1. Vous effectuez une régression logistique pour prédire la probabilité qu'une résidence rencontre un problème de gonflement en fonction de l'âge de la maison et de l'IPPG du remblai sous la dalle de béton. L'IPPG est calculé à partir des proportions de différents faciès dans le remblai. Cette détermination comporte une part de subjectivité et est donc sujette à d'importantes erreurs.

10 *Discutez de l'influence de ces erreurs sur les probabilités prédites en comparaison des probabilités que l'on aurait si l'IPPG était déterminé exactement, et ce, en fonction de la valeur (faible ou forte) de l'IPPG.*

---

2. Vous effectuez un stage au concentrateur d'une mine de Zn et vous êtes en charge du calibrage d'un analyseur en continu aux rayons X du concentré, le Courier-300. Pour effectuer ce calibrage, un échantillon est analysé à intervalles réguliers par l'analyseur Courier-300 et par analyse chimique conventionnelle. La teneur obtenue à l'analyse chimique est considérée comme la vraie teneur de l'échantillon. Le but est de fournir une équation continuellement mise à jour permettant de corriger la teneur obtenue au Courier-300 de façon à mieux estimer les vraies teneurs et permettre ainsi un meilleur ajustement des procédés de flottation.

On vous fournit les quantités suivantes :

Nombre d'échantillons: 5

	Courier-300	Analyse chimique
Somme des valeurs	241	260
Somme des carrés des valeurs	11700	13618
Somme des produits croisés	12617	

10 *a) Un échantillon de concentré analysé au Courier-300 montre une teneur de 50% Zn. Quelle serait la teneur corrigée? (Utilisez un modèle avec constante)*

10 *b) Est-ce que le modèle de régression établi en a) est significatif? Faites le test requis.*(Aide :

vous pouvez soit i) calculer le coefficient de corrélation simple,en déduire le R^2,…ou ii) utiliser le fait que $Y'Y_p=Y'Xb=Y_p'Y_p=SCR$)

5 *c) Supposons que vous disposiez d'analyses chimiques prises à toutes les quatre heures depuis 1 an. Vous voulez établir un modèle de régression "glissant" basé sur les dernières "n" analyses chimiques de façon à permettre au modèle de s'adapter aux dérives temporelles de l'appareil. Le modèle de régression calculé sert à corriger les teneurs du Courrier-300 jusqu'à la prochaine analyse chimique où le modèle est alors mis à jour en ajoutant cette donnée et en laissant tomber la plus ancienne des « n » données. Expliquez comment vous pourriez vous y prendre pour déterminer un « n » optimal en utilisant l'historique des analyses. (Aide: supposons que j'ai obtenu la teneur au Courrier-300 à la 4e heure après la dernière analyse chimique; un bon modèle devrait me permettre de corriger cette teneur pour prédire le plus précisément possible le résultat de l'analyse à venir)*

---

5 3. Dans un réservoir pétrolier, on mesure l'épaisseur totale de shales (S) rencontrés dans les forages et l'épaisseur totale de carbonates (C). On forme une 3e variable comme étant l'épaisseur totale des 2 types de roches (T=S+C).

Que pouvez-vous dire de la corrélation partielle entre S et C étant donné T fixée?

---

20 4. Le tableau suivant présente diverses sommes des carrées pour la régression (modèle avec constante) de données comportant 22 observations et 3 variables. La colonne de gauche est obtenue avec l'ensemble des données, la colonne de droite est obtenue après avoir retirée la donnée #9 et en utilisant les mêmes variables. La somme des carrés des différences entre les valeurs prédites par les deux modèles vaut 4.

Somme des carrés	Modèle avec toutes les observations	Modèle sans l'observation #9
SCTm	503	457
SCRm	254	231
SCE	249	226

Quelle est l'influence de l'observation #9 sur la régression? Qu'en concluez-vous?

---

20 5- On mesure une teneur sur 2 sites différents. On a prélevé n1 observations sur le 1er site et n2 sur le second.

Expliquez comment vous pourriez utiliser un programme de régression comme « regres » pour tester l'égalité des moyennes des teneurs des 2 sites.

---

20 6- On effectue des relevés de diagraphies en forage à l'aide de 2 sondes pour prédire la porosité (exprimée en %). Celle-ci a été déterminée pour 39 carottes prises le long du forage . La 1ère

sonde mesure 3 signaux et la régression procure un R^2 de 0.8. La 2^e sonde mesure 2 signaux et la régression procure un R^2 de 0.7. Lorsque les 5 signaux des 2 sondes sont utilisés conjointement dans la régression, le R^2 atteint 0.85.

D'un point de vue statistique, vaut-il la peine d'utiliser les 2 sondes? La variance expérimentale de la porosité (s_y^2) atteint 112%^{2}.

Bonne Chance

Denis Marcotte

Corrigé :

1- Selon le graphe de p. 35, on voit que le coefficient b (positif) sera sous-estimé. Aux fortes valeurs de l'IPPG, la valeur prédite de W sera donc inférieure à la valeur que l'on aurait si l'on mesurait l'IPPG parfaitement. Aux faibles valeurs ce sera l'inverse. Il en va de même pour les probabilités qui sont obtenus par transformation (monotone) de W.

2- On calcule :

a) x est la mesure au Courrier, y est l'analyse chimique
moy(x)=241/5=48.2
moy(y)=260/5=52
Sxy=(12617- 5*48.2*52)=85
Sx^2=11700-5*48.2^2=83.8
b1=Sxy/Sx^2=85/83.8=1.0143
b0=moy(y)-b1*moy(x)=52-1.0143*48.2=3.110

La valeur corrigée sera donc: 3.11+1.0143*50=53.8%

b) SCR=Yp'*Yp=Y'*Yp=Y'Xb=260*3.1107+12617*1.0143=13606.2
SCM=n*$moy(y)^2$=5*52^2=13520
SCRm=13606.2-13520=86.2
SCTm=13618-13520=98
SCE=98-86.2=11.8
R2=86.2/98=0.88

Autre approche :
r=Sxy/(Sx*Sy)=85/$(98*83.8)^{0.5}$=0.93796
R^2=0.88
SCRm=0.88*SCTm=0.88*98=86

Le reste est identique.

Test (86/1)/(12/3)=86/4=21.5 >F(1,3 ;.05)=10.1 La régression est significative.

c) On fait varier n de 2 à … On détermine les modèles de régression pour chauqe intervalle glissant de n valeurs. Utilisant la n+1 ième valeur au Courrier-300, on effectue la prédiction de l'analyse chimique que l'on compare avec la véritable analyse chimique. Le modèle le plus performant est celui qui fournira les moyennes d'erreur le plus près de 0, et la somme des erreurs au carré ou en valeur absolue la plus faible possible.

3- La corrélation partielle sera –1. En fixant l'épaisseur totale, si l'épaisseur de carbonate augmente d'une valeur « t » il faut que l'épaisseur de shale diminue d'une même valeur.

4- Ici n=22,p=3. CME=SCE(avec toutes les données)/(n-p-1)=249/18=13.83. L'influence est donc 4/(4*13.83)=0.07

L'influence de cette observation n'est pas anormalement forte, il n'y a pas lieu de s'inquiéter.

5- Il suffit de faire un test d'ajout. Le modèle réduit est le modèle avec seulement la constante b0. On définit une variable indicatrice prenant la valeur 1 si l'observation vient du site 1 et 0 sinon (ou l'inverse). On effectue la régression avec cette variable indicatrice. Si l'ajout est significatif c'est que les 2 sites n'ont pas la même moyenne.

6- On effectue la régression modèle complet modèle réduit. La meilleure sonde au départ est la 1ère puisque le R2 est 0.8.
SCTm=112*38=4256
SCRm(r)=0.8*4256=3404.8
SCRm(c)=0.85*4256=3617.6

Test d'ajout : [(3617.6-3404)/2]/[(4256-3617.6)/33]=5.52
Le F (2,33 ;5%) vaut 3.32, on considère que l'ajout est significatif.